N. I. Danilina, N. S. Dubrovskaya,
O. P. Kvasha, G. L. Smirnov

COMPUTATIONAL MATHEMATICS

MIR PUBLISHERS MOSCOW

COMPUTATIONAL MATHEMATICS

Н. И. Данилина, Н. С. Дубровская, О. П. Кваша,
Г. Л. Смирнов

Вычислительная математика

Издательство «Высшая школа» Москва

N. I. Danilina, N. S. Dubrovskaya,
O. P. Kvasha, G. L. Smirnov

COMPUTATIONAL MATHEMATICS

Mir Publishers Moscow

Translated from Russian
by Irene Aleksanova

На английском языке

Printed in the Union of Soviet Socialist Republics

Preface

The rapid development of computer engineering in recent times has led to an expansion of application of mathematics. Quantitative methods have been introduced into practically every sphere of human activity. The use of computers in the economy requires skilled specialists who have a command of the methods of computational mathematics.

Computational mathematics is one of the principal disciplines necessary for the preparation of specialists for various branches of economy. By studying it students acquire theoretical knowledge and practical skill to solve various applied problems with the aid of mathematical models and numerical methods that are realized on a computer.

This study aid assumes that the reader is aware of the elementary concepts of higher mathematics, i.e. continuity, the derivative and the integral. It covers three large divisions of mathematics: "Algebraic Methods" (Ch. 2-6), "Numerical Methods of Analysis" (Ch. 1, 7, 8) and "Numerical Methods of Solving Differential Equations" (Ch. 9, 10).

The theoretical material presented is illustrated by numerous examples. Each chapter is concluded by exercises for independent work.

The following designations are used in the book: the signs □ and ■ are used for the beginning and end of the proof of an assertion and the signs △ and ▲ for the beginning and end of the solution of a problem.

We wish to express our gratitude to Assistant Professor N. I. Ionkin and L. V. Matveeva who reviewed the manuscript and made valuable remarks which thus improved the text.

Authors

Contents

Preface 5
Introduction 11
Chapter 1. Elementary Theory of Errors 14
1.1. Exact and Approximate Numbers. Sources and Classification of Errors 14
1.2. Decimal Notation and Rounding Off Numbers 16
1.3. Absolute and Relative Errors 17
1.4. Valid Significant Digits 20
1.5. The Connection Between the Number of Valid Digits and the Error of the Number 22
1.6. The Errors of a Sum and a Difference . . . 23
1.7. The Error of a Product. The Number of Valid Digits in a Product 27
1.8. The Error of a Quotient. The Number of Valid Digits of a Quotient 32
1.9. The Errors of a Power and a Root 35
1.10. The Rules of Calculating Digits 36
Exercises 38
Chapter 2. Matrix Algebra and Some Data from the Theory of Linear Vector Spaces 39
2.1. Matrices and Vectors. Principal Operations Involving Matrices and Vectors 39
2.2. Transpose of a Matrix 46
2.3. The Determinant of a Matrix. The Properties of the Determinant and the Rules of Its Calculation 48
2.4. The Inverse Matrix 57
2.5. Solving Matrix Equations 63
2.6. Triangular Matrices. Expansion of a Matrix in a Product of Two Triangular Matrices 66
2.7. Inversion of a Matrix by Expanding It in a Product of Two Triangular Matrices . . . 72
2.8. Step Matrices and Operations Involving Them 78
2.9. Inversion of a Matrix by Partitioning It into Blocks 82

2.10. The Absolute Value and the Norm of a Matrix 89
2.11. The Rank of a Matrix and the Methods of Its Calculation 91
2.12. The Concept of a Linear (Vector) Space. The Linear Dependence of Vectors 95
2.13. The Basis of Space 98
2.14. The Transformation of the Coordinates of a Vector upon a Change in the Basis . . . 101
Exercises 105

Chapter 3. Solving Systems of Linear Equations 108
3.1. Systems of Linear Equations 108
3.2. The Kronecker-Capelli Theorem 109
3.3. Cramer's Rule for *n* Linear Equations in *n* Unknowns 111
3.4. Solving Arbitrary Systems of Linear Equations 114
3.5. Homogeneous Systems of Linear Equations 118
3.6. Basic Elimination Procedure 120
3.7. Solving Systems of Linear Equations by the Gauss Elimination Method 124
3.8. Calculating Determinants by the Gauss Elimination Method 136
3.9. The Gaussian Elimination for Inversion of a Matrix 138
3.10. Cholesky's Method 142
3.11. The Iterative Method (the Method of Successive Approximations) 148
3.12. The Conditions for Convergence of an Iterative Process 153
3.13. Estimation of the Error of the Approximate Process of the Iterative Method . . 154
3.14. Seidel's Method. The Conditions for Convergence of Seidel's Process 156
3.15. Estimation of the Errors of Seidel's Process 159
3.16. Reducing a System of Linear Equations to a Form Convenient for Iterations 160
Exercises 162

Chapter 4. Calculating the Values of Elementary Functions 165
4.1. Calculating the Values of Algebraic Polynomials 165
4.2. Calculating the Values of Analytic Functions 170
4.3. The Iterative Method of Calculating the Value of a Function 176
Exercises 178

Chapter 5. Methods of Solving Nonlinear Equations 179
5.1. Algebraic and Transcendental Equations . . 179
5.2. Separating Roots 184

5.3. Computing Roots with a Specified Accuracy. Trial and Error Method 192
5.4. Method of Chords 197
5.5. Newton's Method of Approximation . . . 203
5.6. The Combination of the Method of Chords and Newton's Method 207
5.7. The Iterative Method 214
5.8. General Properties of Algebraic Equations. Determining the Number of Real Roots of an Algebraic Equation 223
5.9. Finding the Domains of Existence of the Roots of an Algebraic Equation 228
5.10. Horner's Method of Approximating Real Roots of an Algebraic Equation 231
Exercises 235
Chapter 6. The Eigenvalues and Eigenvectors of a Matrix 237
6.1. The Characteristic Polynomial 237
6.2. The Method of Direct Expansion 241
6.3. Krylov's Method of Expansion of a Characteristic Determinant 245
6.4. Using Krylov's Method for Calculation of Eigenvectors 253
6.5. The Leverrier-Faddeev Method 254
6.6. Using the Leverrier-Faddeev Method for Calculation of Eigenvectors 258
6.7. Danilevsky's Method 260
6.8. Using Danilevsky's Method for Calculation of Eigenvectors 274
6.9. Using Iterative Methods to Find the First Eigenvalue of a Matrix 277
6.10. Determining the Successive Eigenvalues and the Corresponding Eigenvectors 279
Exercises 281
Chapter 7. Interpolation and Extrapolation 283
7.1. The Function and the Methods of Its Representation 283
7.2. Mathematical Tables 285
7.3. The Approximation Theory 290
7.4. Interpolation by Polynomials 294
7.5. The Error of Interpolation Processes . . . 297
7.6. Lagrange's Interpolating Polynomial . . . 303
7.7. Finite Differences 308
7.8. Stirling and Bessel Interpolating Polynomials 318
7.9. Newton's First and Second Interpolating Polynomials 324
7.10. Divided Differences 329
7.11. Newton's Interpolating Polynomial for an Arbitrary Net of Nodes 334
7.12. Practical Interpolation in Tables 336

7.13. Aitken's Iterated Interpolation 339
7.14. "Optimal-Interval" Interpolation 341
7.15. Interpolation with Multiple Nodes 345
7.16. Mathematical Apparatus of Trigonometric Interpolation 347
7.17. Trigonometric Interpolation 357
7.18. Numerical Methods of Determining the Fourier Coefficients 362
7.19. Backward Interpolation 366
Exercises 371

Chapter 8. Numerical Differentiation and Integration 374
8.1. Statement of a Problem and the Basic Formulas for Numerical Differentiation . . . 374
8.2. Peculiarities of Numerical Differentiation 376
8.3. Statement of a Problem of Numerical Integration 379
8.4. Basic Quadrature Formulas 382
8.5. Newton-Cotes Quadrature Formulas 396
8.6. Quadrature Formulas of the Highest Algebraic Degree of Accuracy 407
8.7. Compounded Quadrature Formulas 412
Exercises 421

Chapter 9. Approximate Solution of Ordinary Differential Equations 423
9.1. Differential Equations 423
9.2. The Method of Successive Approximations (Picard's Method) 426
9.3. Integrating Differential Equations by Means of Power Series 430
9.4. Numerical Integration of Differential Equations. Euler's Method 434
9.5. Modifications of Euler's Method 439
9.6. The Runge-Kutta Method 444
9.7. Adams' Extrapolation Method 452
9.8. Milne's Method 459
9.9. The Notion of the Boundary Value Problem for Ordinary Differential Equations 465
9.10. The Method of Finite Differences for Second-Order Linear Differential Equations . . . 467
Exercises 469

Chapter 10. Approximate Methods of Solution of Partial Differential Equations 471
10.1. Classification of the Second-Order Differential Equations 471
10.2. Classification of Boundary Value Problems 473
10.3. Statement of the Simplest Boundary-Value Problems 477
10.4. The Method of Finite Differences. The Principal Concepts 483

10.5. Difference Schemes for Solving the Equation of Heat Conduction 494
10.6. Difference Schemes for Solving the Equation of Oscillation of a String 498
Exercises 500
Answers to Exercises 502
Index 507

Introduction

Before beginning the exposition of the material, we shall briefly characterize computational mathematics, for which purpose we shall answer the following three questions:

(1) What is computational mathematics?

(2) What are the distinctive features of computational mathematics which allow it to be a special division of mathematics?

(3) What is the significance of computational mathematics for the economy?

1. The term "computational mathematics" means now a division of mathematics which studies problems connected with the use of computers.

We can distinguish three trends in computational mathematics. The first trend is connected with the use of computers in various fields of research and applications and includes, in particular, numerical solution of various mathematical problems. The second trend is connected with the elaboration of new numerical methods and algorithms and perfecting the old ones. The third trend is connected with the problems of the interaction of man and computer.

This book is devoted to the first trend, namely, the use of numerical methods when solving applied problems.

The foundation on which computational mathematics is constructed is composed of various computing facilities, computers first of all, whose rapid development is the most characteristic feature of the technological progress today. Thus, during the last thirty years, the speed of computation increased from one operation per second (with the use of a slide rule) to 3 000 000 operations per second, i.e. $3 \cdot 10^6$ times. It is appropriate to recall that from the time when a steam engine was invented the speed of travel increased from 13 km per hour (the speed of a horse) to 40 000 km per hour (the speed of a cosmic vehicle), i.e. only $3 \cdot 10^3$ times.

2. We can use the following examples to illustrate the characteristic features of computational mathematics which distinguish it from pure mathematics.

From the point of view of a "pure" mathematician, to solve a problem is to prove the existence of its solution and show a process which leads to a solution. For a programmer, the time of obtaining a solution, i.e. the rate of convergence of the process, is often a more important factor. Thus, it is known that a solution of n simultaneous algebraic equations can be theoretically obtained for any specified n as a result of a finite number of operations, say, with the aid of the method of Cramer or Gauss. Therefore, from the viewpoint of a "pure" mathematician, a problem of this kind is considered to be solved. However, when these methods are used in practical applications, two difficulties, which can not always be overcome, are often encountered. The first difficulty is that for a sufficiently large n the number of operations, although finite, is so large that it is impossible to carry out all of them even with the use of the most powerful computers. Thus, to solve a system of n equations by Cramer's method, we must perform $n \cdot n!$ operations, and this constitutes $4.6 \cdot 10^{19}$ operations for $n = 20$. Then, with the rate of $3 \cdot 10^6$ operations per second a computer must operate continuously for half a million years. Gauss' method proves to be more efficient. With the use of this method, the number of operations needed to solve the same problem is of the order of n^3.

However, such a large number of operations generates a second principal difficulty: the errors resulting from all operations accumulate and exert such a great influence on the final result that it often becomes far distant from the true solution.

Nowadays exact methods are usually used for solving systems of equations when their order is not higher than 10^3. Therefore, from the point of view of a programmer, the problem of solving a system whose order is higher than 10^3 is not at all trivial. To solve such a system, iterative methods are used which are approximate but possess a significant advantage of not accumulating computational errors from iteration to iteration. Thus we deal with a seemingly paradoxical situation, but one that is typical of numerical methods, namely, that approximate algorithms are preferred to exact ones.

3. The essential expansion of the fields of application of computational mathematics, including its inculcation into economy, can be explained by the fact that natural phenomena and the phenomena of social life, different in their sense, are often similar in formal structure and can, consequently, be described by the same mathematical models. We can therefore use the same numerical methods to solve the problems described by these models.

Computers are principal factors making for the acceleration of the scientific and technical progress, for the realization of the complex and purposeful programs of solution of the most important scientific and technological problems and for the further increase in the productivity of labour. The development of computers and computational mathematics will make it possible, for instance, to pass from the automatization of the control of technological systems and processes to the automatization of the control of production processes.

Chapter 1

Elementary Theory of Errors

1.1. Exact and Approximate Numbers. Sources and Classification of Errors

In the process of solving a problem, we have to deal with various numbers which may be exact or approximate. Exact numbers give a true value of a number and approximate numbers give a value close to the true one, the degree of closeness being dependent on the error of calculation.

For example, in the assertions "a cube has six faces". "we have five fingers to a hand", "there are 32 students in a class", "there are 582 pages in a book" the numbers 6, 5, 32 and 582 are exact ones. In the assertions "the house is 14.25 m wide", "the radius of the Earth is 6000 km", "the mass of a match box is ten g" the numbers 14.25, 6000 and 10 are approximate.

This is due, first of all, to the imperfection of measuring instruments we use. There are no absolutely exact measuring instruments, each of them has its own accuracy, i.e. admits of a certain error of measurements. In addition, in the second example the approximation of a number is in the very concept of the radius of the Earth. The matter is that, strictly speaking, the Earth is not a sphere and we can speak of its radius only in approximate terms. In the next example, the approximation of the number is also defined by the fact that different boxes may have different masses and the number 10 defines the mass of a certain box.

In other cases, the same number may be exact as well as approximate. Thus, for instance, the number 3 is exact if we speak of the number of sides of a triangle and approximate if we use it instead of the number π when calculating the area of a circle using the formula $S = \pi R^2$.

In practical calculations, we understand the *approximate number a* to be a number which differs but slightly from the exact number A and can be substituted for it in calculations.

The solution of the majority of practical problems with a certain degree of conventionality can be represented as two successive stages: (1) the mathematical description of the problem on hand, (2) the solution of the formulated mathematical problem.

At the first stage, we may encounter two characteristic sources of errors. First, the fact that the processes happening in reality can not always be described by means of mathematics and the simplifications we introduce make it possible to obtain only more or less idealized models. Second, the initial parameters are, as a rule, inexact since they are obtained from an experiment which gives only an approximate result.

Accordingly, the total error of a mathematical model and initial data is considered to be the *error of the initial information*. Having in mind that this error is independent of the second stage of solving the problem, we often call it a *nonremovable error*.

It is, as a rule, unrealizable in practice to obtain an exact solution of a mathematical problem (the second stage) irrespective of whether it is constructed analytically or on a computer. Thus, for instance, we can obtain an exact solution for only a very restricted class of differential equations. Therefore, in practical calculations we usually use the methods of approximation of solutions, numerical first of all.

Such a compulsory replacement of an exact solution by an approximate one generates an *error of the method* or, as it is often called, an *error of approximation*.

Finally, in the process of problem solving, we round off the initial data as well as the intermediate and final results. These errors and the errors arising in the arithmetic operations involving approximate numbers affect, more or less, the result of calculations and form a so-called *rounding error*.

In this connection, when we formulate a problem, we either indicate the accuracy of the solution required, i.e. specify the maximum error permissible in all calculations, or only calculate the total error of the result. Therefore, when dealing with approximate numbers, it is necessary to know how to solve the following problems:

(1) to characterize the exactness of approximate numbers by mathematical means,

(2) to estimate the degree of accuracy of the result when we know the degree of accuracy of the initial data,

(3) to choose initial data with the degree of accuracy which will ensure the specified accuracy of the result,

(4) to construct an optimal computing process in order to obviate the calculations which do not affect the valid digits of the result.

1.2. Decimal Notation and Rounding off Numbers

Every decimal positive number a can be represented as a finite or infinite decimal fraction

$$a = \alpha_1 \cdot 10^m + \alpha_2 \cdot 10^{m-1} + \ldots + \alpha_n \cdot 10^{m-n+1} + \ldots, \quad (1)$$

where α_i are the digits constituting the number ($i = 1, 2, \ldots, n, \ldots$) with $\alpha_1 \neq 0$, and m is the top digit in the number a.

Example 1. Represent the number 1905.0778 in form (1):

$$1905.0778 = 1 \cdot 10^3 + 9 \cdot 10^2 + 0 \cdot 10^1 + 5 \cdot 10^0 + 0 \cdot 10^{-1} + 7 \cdot 10^{-2} + 7 \cdot 10^{-3} + 8 \cdot 10^{-4}.$$

Every unit in the corresponding ith decimal position, reckoning from left to right, has its value 10^{m-i+1} known as the *value of the decimal position*. Thus the value of the first (from the left) decimal position is 10^m, that of the second is 10^{m-1} and so on.

In the example considered, the value of the decimal position containing the digit 9 is $10^{3-2+1} = 100$, of that containing the digit 5 is $10^{3-4+1} = 1$, of that containing the digit 8 is $10^{3-8+1} = 0.0001$.

In practical calculations we often have to round off a number, i.e. to replace it by another number consisting of a smaller number of digits. In that case we retain one or several digits, reckoning from left to right, and discard all the others.

The following *rules of rounding off* are most often used.

1°. *If the discarded digits constitute a number which is larger than half the unit in the last decimal place that remains, then the last digit that is left is strengthened (increased by unity).*

Now if the discarded digits constitute a number which is smaller than half the unit in the last decimal place that remains, then the digits that remain do not change.

2°. *If the discarded digits constitute a number which is equal to half the unit in the last decimal place that remains, then the last digit that is left is strengthened, if it is odd, and is unchanged if it is even.*

This rule is often called a **rule of an even digit.**

Example 2. Round off the following numbers: $A_1 = 12.7852$, $A_2 = 394.261$, $A_3 = 6.265001$, $A_4 = 147.5$, $A_5 = 148.5$ to three digits.

△ In accordance with item 1° of the rules of rounding off, we get $a_1 = 12.8$, $a_2 = 394$, $a_3 = 6.27$ since $0.0852 > 0.5 \cdot 10^{-1}$, $0.261 < 0.5 \cdot 10$, $0.005001 > 0.5 \cdot 10^{-2}$.

In accordance with item 2° of the rules of rounding off, we get $a_4 = 148, a_5 = 148$ since the digit 7 is odd and the digit 8 is even. ▲

In some cases which are more and more often encountered nowadays, use is made of a more simple rule of rounding off. This rule consists in a simple discarding of all digits beginning with a certain decimal place. Using this rule, we would get the following values when rounding off the numbers from the examples considered: $a_1 = 12.7$, $a_2 = 394$, $a_3 = 6.26$, $a_4 = 147$, $a_5 = 148$.

1.3. Absolute and Relative Errors

Let A be an exact number and a, its approximate value. If $a < A$, then we say that the number a is an *approximate value of the number A by defect* and if $a > A$, then it is an *approximate value of A by excess.*

The difference between the exact number A and its approximation a is an *error.*

As a rule, it is impossible to determine the value of the error $A - a$ and even its sign since the exact number A is unknown. Therefore we use the upper bound of the absolute value of the error rather than the error itself.

The *absolute error* of the approximate number a is a quantity Δ_a which satisfies the inequality

$$\Delta_a \geqslant |A - a|. \tag{1}$$

The absolute error is the upper bound of the deviation of the exact number A from its approximation:

$$a - \Delta_a \leqslant A \leqslant a + \Delta_a. \tag{2}$$

Inequality (2) is often written in the form

$$A = a \pm \Delta_a. \tag{3}$$

The number taken as an absolute error must be as small as possible. For instance, measuring the length of a line segment, we have found that the error of measurement does not exceed 0.5 cm, all the more so it does not exceed 1, 2 and 3 cm. Each of these numbers can be taken as the absolute error. However, we must take the smallest of these numbers as the absolute error since the smaller the absolute error, the narrower the interval within which we specify the exact number.

In practical calculations we often use expressions "with an accuracy to 0.01", "with an accuracy to 1 cm" etc. This means that the absolute error is equal to 0.01, 1 cm etc. respectively.

Example 1. We have measured the length of the line segment L with an accuracy to 0.05 cm and obtained $l = 18.4$ cm. The absolute error here $\Delta_l = 0.05$ cm. In accordance with formula (3), we must write $L = 18.4 \pm 0.05$ cm. According to formula (2), the exact value of the length of the segment is within the interval $18.35 \leqslant L \leqslant 18.45$.

Example 2. We have measured the length of the line segment L using a ruler with the value of the division 0.1 cm. We have found that the exact value of L is between 4.6 and 4.7. In this case, we must take $l = 4.65$, i.e. the middle of the interval within which the exact number L is, as the approximate value. The absolute error is, evidently, half the value of the division of the ruler, i.e. $\Delta_l = 0.05$. Thus $L = 4.65 \pm 0.05$ cm.

The absolute error reflects only the quantitative aspect of the error but not the qualitative one, i.e. does not show whether the measurement and calculation were accurate. Indeed, assume that measuring the length and the width of the top of a table with a ruler the value of whose division is 1 cm, we have got the following results (in cm): the width $L_1 = 2 \pm 0.5$ and the length $L_2 = 100 \pm 0.5$. In both measurements the absolute error is the same and constitutes 0.5 cm. It is evident, however, that the second measurement was more accurate than the first. To estimate the quality of calculations or measurements, the concept of a relative error is introduced.

The *relative error* of the approximate number a is the quantity δ_a which satisfies the inequality

$$\delta_a \geqslant \left| \frac{A-a}{a} \right|, \quad a \neq 0. \tag{4}$$

In particular, we can accept

$$\delta_a = \frac{\Delta_a}{|a|}, \quad a \neq 0, \tag{5}$$

as the relative error and represent relation (3) in the form

$$A = a\,(1 \pm \delta_a). \tag{6}$$

Note that a relative error is an abstract number and is often expressed in per cent.

Now we return to the measurements of the length and the width of the top of the table and find their relative errors:

$$\delta_{l_1} = 0.5/2 = 0.25 \text{ or } 25\%, \quad \delta_{l_2} = 0.5/100 = 0.005 \text{ or } 0.5\%.$$

In such cases we say that the measurement of the length of the top of the table has been relatively more accurate (50 times as accurate) than that of its width.

Example 3. An exact number A is in the interval $[23.07, 23.10]$. Find its approximate value, the absolute and the relative error.

△ We assume the middle of the given interval to be its approximate value: $a = 23.085$. The absolute error is half its length: $\Delta_a = 0.015$. We assume $\delta_a = \Delta_a/a = 0.000604\ldots$ to be the relative error. It is customary to round off the value of the error to one or two nonzero digits. Therefore we can set $\delta = 0.07\%$. Note that in problems of this kind the error is usually rounded off to a larger number in order that inequalities (2) and (4) should be satisfied. ▲

Example 4. Determine, in per cent, the relative error of the approximate number $a = 35.148$ if $A = 35.148 \pm 0.00074$.

△ Using formula (5), we have

$$\delta_a = \Delta_a/a = 0.00074/35.148 = 0.000022 \cong 0.003\%. \ ▲$$

Example 5. Determine the absolute error of the approximate number $a = 4.123$ if $\delta_a = 0.01\%$.

△ We write the percentage in the form of a decimal fraction and use formula (5) to find the absolute error. Then we have

$$\Delta_a = |a| \cdot \delta_a = 4.123 \cdot 0.0001 \cong 0.0005,$$

$$A = 4.123 \pm 0.0005. \ ▲$$

Example 6. Find out in which of the two following cases the quality of calculations is higher: $A_1 = 13/19 \cong 0.684$ or $A_2 = \sqrt{52} \cong 7.21$.

△ To find the absolute errors, we take numbers a_1 and a_2 with a larger number of decimal digits: $13/19 \cong 0.68421$, $\sqrt{52} \cong 7.2111\ldots$. We determine the absolute errors by rounding them off to a larger number:

$$\Delta_{a_1} = |0.68421\ldots - 0.684| \cong 0.00022,$$

$$\Delta_{a_2} = |\,7.2111\ldots - 7.21\,| \cong 0.0012.$$

We seek the relative errors:

$$\delta_{a_1} = \Delta_{a_1}/|\,a_1\,| = 0.00022/0.684 \cong 0.00033 \cong 0.04\%,$$
$$\delta_{a_2} = \Delta_{a_2}/|\,a_2\,| = 0.0012/7.21 \cong 0.00017 \cong 0.02\%.$$

In the second case the quality of calculation proved to be higher since $\delta_{a_2} < \delta_{a_1}$. ▲

1.4. Valid Significant Digits

When we solve a problem, we often stipulate that the result must be obtained with an accuracy of 0.1, 0.01 etc. It may seem that the accuracy of calculations is defined by the number of digits after the decimal point. This is not so, however. The accuracy of calculations is defined by the number of digits in the result which enjoy confidence.

The *significant digits* of a number are all its digits, except for zeros, which appear to the left of the first nonzero digit.

Zeros at the end of a number are always significant digits (otherwise they are not written).

Example 1. The numbers 0.001604 and 30.500 have 4 and 5 significant digits respectively.

There are certain subtle points in the representation of integers. Thus, for instance, if we want to show that the last three zeros in the number 400 000 are not significant, we must write this number as two factors: $400 \cdot 10^3$, or $40.0 \cdot 10^4$, or $0.400 \cdot 10^6$. The last form of the notation is known as *normalized* and is preferable. In this case we say that 400 is the *mantissa* of the number and 6 is its *order*.

Recall that every positive decimal number, exact as well as approximate, can be represented as

$$a = \alpha_1 \cdot 10^m + \alpha_2 \cdot 10^{m-1} + \ldots + \alpha_n \cdot 10^{m-n+1} + \ldots$$

The digit α_n of the approximate number a is a *valid significant digit* (or simply a *valid digit*) if there holds an inequality

$$|\,A - a\,| \leqslant 0.5 \cdot 10^{m-n+1}, \qquad (1)$$

i.e. if the absolute value of the difference of the exact number and its approximate value does not exceed half the unit of the decimal digit in which α_n appears.

Since instead of $|A - a|$ we usually consider the absolute error Δ_a, we often replace inequality (1) by an inequality

$$\Delta_a \leqslant 0.5 \cdot 10^{m-n+1} \tag{2}$$

because when this inequality is satisfied, the initial inequality (1) is satisfied too.

On the other hand, if the number n of valid digits of the approximate number a is specified, then we can take

$$\Delta_a = 0.5 \cdot 10^{m-n+1} \tag{3}$$

as the absolute error.

If inequality (2) is not satisfied, then the digit α_n is said to be *doubtful*. It is evident that if the digit α_n is valid, then all the preceding digits, to the left of it, are also valid.

Example 2. We have obtained the number $a = 23.10$ by rounding off an exact number. How many valid digits are there in the number a?

△ When a number is rounded off according to the rule of an even digit, the absolute error cannot exceed half the unit in the last decimal place that remains. This means that all the remaining digits in the rounded-off number are valid. In this case, evidently, all the four digits are valid and the error $\Delta_a = 0.005$. ▲

Example 3. The number $a = 23.071937$ contains five valid digits. Find its absolute error.

△ We use formula (3). Here $m = 1$, $n = 5$, and so we can take $\Delta_a = 0.5 \cdot 10^{1-5+1} = 0.0005$ as the absolute error. ▲

Example 4. The absolute error of the number $a = 705.1978$ is $\Delta_a = 0.3$. Find out which of the digits of the number a are valid and round off the number a leaving only valid digits.

△ We use formula (2). Here $m = 2$, $\Delta_a = 0.3$ and n must be found from the inequality $0.3 \leqslant 0.5 \cdot 10^{3-n}$. A direct verification shows that the greatest n which satisfies this inequality is equal to 3 and the digit 5 is valid: $0.3 < 0.5 \cdot 10^{2-3+1}$, and the digit 1 is doubtful: $0.3 > 0.5 \cdot 10^{2-4+1}$.

Consequently, the number $a = 705.1978$ has three valid digits. We round it off to three digits: $a_1 = 705$. Then the total error is equal to the sum of the initial error and the rounding error: $\Delta_a = 0.3 + 0.2 = 0.5$, and so we can write $A = 705 \pm 0.5$ ▲

In mathematical tables all significant digits are valid, as a rule. Thus in the well-known tables of logarithms the values of the sine are given with an absolute error of $0.5 \cdot 10^{-4}$.

In recent years, the concept of valid significant digits is more and more often used in the broad sense. This concept is connected with the simplest rule of rounding off which we mentioned in 1.2.

The digit α_n of the approximate number

$$a = \alpha_1 \cdot 10^m + \alpha_2 \cdot 10^{m-1} + \ldots + \alpha_n \cdot 10^{m-n+1} + \ldots$$

is a *valid significant digit in the broad sense* if there holds an inequality

$$\Delta_a \leqslant 1 \cdot 10^{m-n+1}, \tag{4}$$

i.e. if the absolute error of the number a does not exceed the unit of the decimal place in which α_n appears.

1.5. The Connection Between the Number of Valid Digits and the Error of the Number

As follows from the definition of a valid significant digit, the number of the valid digits of an approximate number is defined by the inequality

$$|A - a| \leqslant 0.5 \cdot 10^{m-n+1}. \tag{1}$$

Dividing both sides of inequality (1) by $|a|$, we get

$$\left|\frac{A-a}{a}\right| \leqslant \frac{0.5 \cdot 10^{m-n+1}}{|\alpha_1 \cdot 10^m + \alpha_2 \cdot 10^{m-1} + \ldots + \alpha_n \cdot 10^{m-n+1} + \ldots|}$$

$$\leqslant \frac{0.5 \cdot 10^{m-n+1}}{\alpha_1 \cdot 10^m} = \frac{0.5}{\alpha_1 \cdot 10^{n-1}}. \tag{2}$$

Thus, if the digit α_n of the approximate number a is valid, then we can assume

$$\delta_a = \frac{0.5}{\alpha_1 \cdot 10^{n-1}} \tag{3}$$

to be the relative error.

On the other hand, for the digit α_n of the approximate number a to be valid, it is necessary that the inequality

$$\delta_a \leqslant \frac{0.5}{(\alpha_1 + 1) \cdot 10^{n-1}} \tag{4}$$

hold true since in this case inequalities (2) and (1) are satisfied.

When we mean valid significant digits in the broad sense, we can get a similar formula

$$\delta_a = \frac{1}{\alpha_1 \cdot 10^{n-1}}. \tag{5}$$

Example 1. What is the relative error of the approximate number $a = 4.176$ if all its digits are valid?

△ Since all the four digits of the number 4.176 are valid, we use formula (3) to find the relative error:

$$\delta_a = \frac{0.5}{\alpha_1 \cdot 10^{n-1}} = \frac{1}{2 \cdot 4 \cdot 10^3} \cong 0.00013 = 0.013\%.$$

Note that the relative error of the number a can be found from the formula $\delta_a = \Delta_a / | a |$. Since in the given number a all digits are valid, we have $\Delta_a = 0.0005$. Then

$$\delta_a = 0.0005/4.176 \cong 0.00012 = 0.012\%.$$

We can see that the difference is not large, but applying formula (3), we somewhat simplify the calculations. ▲

Example 2. What is the relative error of the number $a = 14.278$ if all its digits are valid in the broad sense?

△ Since all the five digits of the number are valid in the broad sense, we can use formula (5) to obtain

$$\delta_a = \frac{1}{\alpha_1 \cdot 10^{n-1}} = \frac{1}{1 \cdot 10^4} = 0.0001 = 0.01\%. \ ▲$$

Example 3. How many decimal digits must be taken in the number $\sqrt{18}$ for the error not to exceed 0.1%?

△ Here $A = \sqrt{18} \cong 4, \ldots$; $\delta_a \leqslant 0.1\%$, i.e. $\delta_a \leqslant 0.001$. We have $\delta_a = \frac{1}{2 \cdot 4 \cdot 10^{n-1}} \leqslant 0.001$, whence $125 \leqslant 10^{n-1}$; $1.25 \times 10^2 \leqslant 10^{n-1}$; $\log 1.25 + 2 \leqslant n - 1$; $n \geqslant 3 + \log 1.25$, i.e. $n \geqslant 4$. ▲

1.6. The Errors of a Sum and a Difference

Consider the exact numbers $A_1, A_2, \ldots, A_n$ and their approximations $a_1, a_2, \ldots a_n$. Let $A = \sum_{i=1}^{n} A_i$ be the sum of all exact numbers and $a = \sum_{i=1}^{n} a_i$ be the sum of their approximations. We pose the following problem: being given the absolute errors $\Delta_{a_1}, \Delta_{a_2}, \ldots, \Delta_{a_n}$ of all approximate numbers, evaluate the absolute error of their sum a. We set up a difference

$$A - a = (A_1 - a_1) + (A_2 - a_2) + \ldots + (A_n - a_n).$$

Passing to the absolute values on the right-hand and left-hand sides of the relation and using the property of absolute values, we obtain

$$|A-a| \leqslant |A_1-a_1| + |A_2-a_2| + \ldots + |A_n-a_n|.$$

Consequently,

$$|A-a| \leqslant \Delta_{a_1} + \Delta_{a_2} + \ldots + \Delta_{a_n} \tag{1}$$

and we can take the sum of the absolute errors of the terms

$$\Delta_a = \Delta_{a_1} + \Delta_{a_2} + \ldots + \Delta_{a_n}, \tag{2}$$

as the absolute error of the approximate number a, i.e. the sum of the approximate numbers $a_1, a_2, \ldots, a_n$.

It follows from the last formula that the absolute error of the algebraic sum must, in general, be not smaller than the absolute error of the least exact term. Therefore, to obviate excess calculations, we should not leave unnecessary digits in the more exact terms either.

When adding up numbers of different absolute accuracy, we usually do the following:

(1) isolate a number (or numbers) of the least accuracy (i.e. a number which has the greatest absolute error),

(2) round off more exact numbers so as to retain in them one digit more than in the isolated number (i.e. retain one reserve digit),

(3) perform addition taking into account all the retained digits,

(4) round off the result obtained by discarding one digit.

Remark. When the number of terms is large ($n > 10$), the evaluation of the error of the sum by formula (2) proves to be too high since a partial compensation of the errors of different signs usually occurs. If all the terms are rounded off to the mth decimal place, i.e. their errors are evaluated by the quantity $0.5 \cdot 10^{-m}$, then the statistical evaluation of the absolute error of the sum can be found from the following formula:

$$\Delta_a = \sqrt{n} \cdot 0.5 \cdot 10^{-m}. \tag{3}$$

Example 1. Add up the approximate numbers $a = 0.1732 + 17.45 + 0.000333 + 204.4 + 7.25 + 144.2 + 0.0112 + 0.634 + 0.0771$ in each of which all the written digits are valid.

△ We choose the least exact numbers (those possessing the greatest absolute error). There are two numbers of this kind, 204.4

and 144.2. The error of each of them is 0.05. We round off the other numbers, leaving one (reserve) sign more, and add up all the numbers:

$$
\begin{array}{r r}
 & 0.17 \\
 & 17.45 \\
 & 0.00 \\
 & 204.4 \\
+ & 7.25 \\
 & 144.2 \\
 & 0.01 \\
 & 0.63 \\
 & 0.08 \\
\hline
 & 374.19
\end{array}
$$

We round off the sum obtained discarding one digit: 374.2.

We estimate the accuracy of the result. The absolute error of the sum consists of two terms:

(1) the initial error, i.e. the sum of the errors of the least exact numbers and the rounding errors of the other numbers: $0.05 \cdot 2 + 0.005 \cdot 7 \cong 0.14$,

(2) the error of the rounding off the result: 0.01.

Thus the absolute error of the sum is 0.15 and we must write the result in the form $A = 374.2 \pm 0.15$. Another form of notation, $A = 374.2 \pm 0.2$ is also possible. ▲

We do the same in the case when one or several approximate numbers are negative.

Example 2. Find the difference of the approximate numbers $a = a_1 - a_2$ and evaluate the absolute and the relative error of the result if $A_1 = 17.5 \pm 0.02$ and $A_2 = 45.6 \pm 0.03$.

△ We find that $a = a_1 - a_2 = 17.5 - 45.6 = -28.1$; $\Delta_a = \Delta_{a_1} + \Delta_{a_2} = 0.02 + 0.03 = 0.05$. Thus $A = -28.1 \pm 0.05$. We find the relative error: $\delta_a = 0.05/|-28.1| \cong 0.002 = 0.2\%$. ▲

We can show that if the absolute error of the sum of approximate numbers can be found from formula (2) and the relative error of the sum $\delta_a = \Delta_a/|a|$, then

$$\delta_{\min} \leqslant \delta_a \leqslant \delta_{\max}.$$

Example 3. Evaluate the relative error of the sum of the numbers in Example 1 and compare it to the relative errors of the terms.

△ We find the relative error of the sum:

$$\delta_a = 0.2/374.2 = 0.0006 = 0.06\%.$$

The relative errors of the terms are

$$
\begin{aligned}
&\delta_{a_1} = 0.005/0.17 = 3\%, \quad &&\delta_{a_2} = 0.005/7.45 = 0.03\%, \\
&\delta_{a_4} = 0.05/204.4 = 0.03\%, \quad &&\delta_{a_5} = 0.005/7.25 = 0.07\%, \\
&\delta_{a_6} = 0.05/144.2 = 0.04\%, \quad &&\delta_{a_7} = 0.005/0.01 = 50\%, \\
&\delta_{a_8} = 0.005/0.63 = 0.8\%, \quad &&\delta_{a_9} = 0.005/0.08 = 7\%.
\end{aligned}
$$

Thus $\delta_{\min} = 0.03\%$, $\delta_{\max} = 50\%$, $\delta_a = 0.06\%$, i.e. the relative error of the sum is between the least and the greatest relative error of the terms. ▲

Note that when subtraction concerns close numbers, a situation may occur which is known as a *loss of accuracy*. Let $x > 0$, $y > 0$ and $a = x - y$. Then

$$\delta_a = \frac{\Delta_a}{|a|} = \frac{\Delta_x + \Delta_y}{|x-y|}.$$

Thus if the numbers x and y differ but little from each other, then even for small errors Δ_x and Δ_y, the value of the relative error of the difference may turn to be considerable.

Example 4. **Let $x = 5.125$, $y = 5.135$. Here $\Delta_x = 0.0005$, $\Delta_y = 0.0005$, $\delta_x \cong \delta_y \cong 0.01\%$. The relative error of the difference $a = x - y$ is**

$$\delta_a = \frac{0.0005 + 0.0005}{0.01} \cdot 100 = 10\%.$$

Evidently, when subtraction concerns two close numbers, a considerable loss of accuracy may occur. To obviate this, we must change the subtraction procedure so that small differences of the quantities are calculated directly.

Example 5. **Find the difference $A = \sqrt{6.27} - \sqrt{6.26}$ and evaluate the relative error of the result.**

△ Let $A_1 = \sqrt{6.27} \cong 2.504$, $\Delta_{a_1} = 0.0005$, $A_2 = \sqrt{6.26} \cong 2.502$, $\Delta_{a_2} = 0.0005$. Then $a = 2.504 - 2.502 = 0.2 \cdot 10^{-2}$, $\Delta_a = 0.0005 + 0.0005 = 0.001$, whence

$$\delta_a = \frac{0.1 \cdot 10^{-2}}{0.2 \cdot 10^{-2}} = 0.5 = 50\%.$$

However, changing the scheme of calculation, we can get a better evaluation of the relative error:

$$A = \sqrt{6.27} - \sqrt{6.26} = \frac{(\sqrt{6.27} - \sqrt{6.26})(\sqrt{6.27} + \sqrt{6.26})}{\sqrt{6.27} + \sqrt{6.26}}$$

$$= \frac{6.27 - 6.26}{\sqrt{6.27} + \sqrt{6.26}} = \frac{0.01}{\sqrt{6.27} + \sqrt{6.26}} \cong 0.2 \cdot 10^{-2} = a,$$

$$\delta_a = \frac{\Delta_{a_1} + \Delta_{a_2}}{a_1 + a_2} = \frac{0.001}{5} \cong 0.2 \cdot 10^{-3} = 0.02\%.$$

Thus, when calculating a_1 and a_2 with the same four valid digits, we have got an essentially better result in the sense of a relative error. ▲

Example 6. Calculate the value of the function $y = 1 - \cos x$ for the following values of the argument: (1) $x_1 = 80°$, (2) $x_2 = 1°$. Calculate the absolute and the relative error of the result.

△ (1) From the four-digit logarithmic tables we find that $\cos 80° \cong 0.1736$ and, since all the digits of this number are valid, we have $\Delta_{0.1736} = 0.00005$. Then $y_1 = 1 - 0.1736 = 0.8264$ and $\Delta_{y_1} = 0.00005$ (from an exact number equal to unity we subtract an approximate number with an absolute error not exceeding 0.00005). Consequently,

$$\delta_{y_1} = 0.00005/0.8264 = 0.00006 = 0.006\%.$$

(2) We have $\cos 1° \cong 0.9998$, $\Delta_{0.9998} = 0.00005$, $y_2 = 1 - 0.9998 = 0.0002$, $\Delta_{y_2} = 0.00005$, hence

$$\delta_{y_2} = 0.00005/0.0002 = 0.25 = 25\%.$$

We can see from the examples presented that for small values of the argument a direct calculation by the formula $y = 1 - \cos x$ yields a relative error of the order of 25%. For $x = 80°$ the relative error is only 0.006%.

We change the calculation procedure and use a formula $y = 1 - \cos x = 2\sin^2(x/2)$ to calculate the values of the function $y = 1 - \cos x$ for small values of the argument. We designate $a = \sin 0°30' \cong 0.0087$. Then $\Delta_a = 0.00005$, $\delta_a = 0.5/87 = 0.58\%$. Furthermore,

$$y_2 = 2\cdot 0.0087^2 \doteq 0.000151,$$
$$\delta y_2 = 2\cdot 0.0058 \doteq 1.2\%$$

(see 1.7 below). The final result is

$$\Delta_{y_2} = y_2\delta_{y_2} = 0.000151\cdot 0.012 = 0.000002$$

(whereas earlier we had $\Delta_{y_2} = 0.00005$). Thus a simple transformation of the computing formula has allowed us to get a more accurate result from the same initial data. ▲

It is not always possible, however, to transform the computing procedure. Therefore, when close numbers are subtracted, they must be taken with a sufficient number of reserve valid digits (when it is possibbe). If it is known that the first m significant digits may be lost and we must get a result with n valid significant digits, we must take the initial data with $m + n$ valid significant digits as was done in Example 5.

1.7. The Error of a Product. The Number of Valid Digits in a Product

The error of a product. Let us consider two exact numbers A_1 and A_2 and their approximate values a_1 and a_2. Let $A = A_1A_2$ and $a = a_1a_2$. We pose the following problem: being given relative errors δ_{a_1} and δ_{a_2}, evaluate the relative error of the product δ_a.

We represent the exact values A_1 and A_2 in the form

$$A_1 = a_1 + \Delta_1, A_2 = a_2 + \Delta_2, \tag{1}$$

where the unknowns Δ_1 and Δ_2 satisfy the inequalities

$$|\Delta_1| \leqslant |a_1| \delta_{a_1}, |\Delta_2| \leqslant |a_2| \delta_{a_2}. \tag{2}$$

Multiplying the right-hand and left-hand sides of relations (1), we obtain

$$A_1A_2 = a_1a_2 + \Delta_1a_2 + \Delta_2a_1 + \Delta_1\Delta_2.$$

Passing to absolute values on the right-hand and left-hand sides of this relation and using the properties of absolute values, we find that

$$|A_1A_2 - a_1a_2| \leqslant |\Delta_2a_1| + |\Delta_1a_2| + |\Delta_1\Delta_2|. \tag{3}$$

Since the last term on the right-hand side is small, we discard it and divide the right-hand and left-hand sides of the inequality by $|a| = |a_1a_2|$. Then, taking relation (2) into account, we obtain

$$\left|\frac{A-a}{a}\right| \leqslant \delta_{a_1} + \delta_{a_2}. \tag{4}$$

It follows from the relation obtained that we can take the sum of the relative errors of the factors

$$\delta_a = \delta_{a_1} + \delta_{a_2} \tag{5}$$

as the relative error of the product $a = a_1a_2$.

Inequality (5) can be easily extended to the product of several factors so that if $A = A_1A_2 \ldots A_n$ and $a = a_1a_2 \ldots a_n$, then we can assume that

$$\delta_a = \delta_{a_1} + \delta_{a_2} + \ldots + \delta_{a_n}. \tag{6}$$

In the case when all factors, except for one, are exact numbers, it follows from formula (6) that the relative error of the product coincides with the relative error of the approximate factor. Thus, when only the value of the factor a_i is an approximate number, then

$$\delta_a = \delta_{a_i}. \tag{7}$$

Remark. When the approximate number a is multiplied by the exact factor k, the relative error of the product is equal to the relative error of the approximate number a and the absolute error is $|k|$ times as large as the absolute error of the approximate number.

Indeed, let $a = ka_1$, where k is an exact factor different from zero. Then, according to formula (7), we have $\delta_a = \delta_{a_1}$, or

$$\Delta_a = |a|\,\delta_a = |a|\,\delta_{a_1} = |ka_1|\,\frac{\Delta_{a_1}}{|a_1|} = |k|\,\Delta_{a_1},$$

i.e.

$$\Delta_a = |k|\,\Delta_{a_1}. \tag{8}$$

Knowing the relative error δ_a of the product a, we can find its absolute error using the formula $\Delta_a = |a|\,\delta_a$.

If the relative error of the product of approximate numbers can be found from formula (6), then, when multiplying numbers with different relative errors, we may not retain the extra digits in the numbers with the smaller relative errors. We usually do the following:

(1) isolate a number with the least number of valid digits,

(2) round off the remaining factors so that they would contain one significant digit more than there are valid significant digits in the isolated number,

(3) retain as many significant digits in the product as there are valid significant digits in the least exact factor (the isolated number).

Example 1. Find the product of the approximate numbers $x_1 = 3.6$ and $x_2 = 84.489$ all of whose digits are valid.

△ In the first number there are two valid significant digits and in the second there are five. Therefore we round off the second number to three significant digits. After rounding-off we have $x_1 = 3.6$, $x_2 = 84.5$. Hence

$$x_1x_2 = 3.6\cdot 84.5 = 304.20 \approx 3.0\cdot 10^2.$$

There are two significant digits in the result, i.e. as many as there were significant digits in the factor with the least number of valid significant digits. ▲

Example 2. Find the product of the approximate numbers $x_1 = 12.4$ and $x_2 = 65.54$ and the number of valid digits in it if all the written digits in the factors are valid.

△ The first number contains three valid significant digits and the second number contains four. We can multiply the numbers without previous rounding-off, i.e. $x_1x_2 = 12.4 \cdot 65.54 = 812.696$. We must retain three significant digits since the least exact factor has the same number of valid significant digits. Thus $a = 813$. Let us calculate the error:

$$\delta_a = \delta_{x_1} + \delta_{x_2} = \frac{0.05}{12.4} + \frac{0.005}{65.54} = 0.0041.$$

Then $\Delta_a = 813 \cdot 0.0041 \approxeq 3.4$. This means that the product has two valid digits and should be written as $A = 813 \pm 4$. ▲

The number of valid digits in the product. Consider a product of k factors ($k \leqslant 10$) $a = a_1a_2 \ldots a_k$, where $a_i \neq 0$. Each factor contains no less than n valid digits ($n > 1$).

Assume that each of the factors has the form

$$a_i = \alpha_i \cdot 10^{l_i} + \beta_i \cdot 10^{l_i - 1} + \gamma_i \cdot 10^{l_i - 2} + \ldots$$
$$(i = 1, 2, \ldots, k), \tag{9}$$

where α_i are the first significant digits of the approximate factors written in the decimal notation.

For the relative error of the approximate number which has n valid digits we use the formula

$$\delta_{a_i} = \frac{0.5}{\alpha_i \cdot 10^{n-1}} \quad (i = 1, 2, \ldots, k).$$

Then the relative error of the product of k approximate numbers, each of which has n valid significant digits, is

$$\delta_a = \delta_{a_1} + \delta_{a_2} + \ldots + \delta_{a_k} = \frac{0.5}{10^{n-1}} \left(\frac{1}{\alpha_1} + \frac{1}{\alpha_2} + \ldots + \frac{1}{\alpha_k} \right). \tag{10}$$

Taking into account that the number of factors is not larger than 10 ($k \leqslant 10$), we obtain

$$\frac{1}{\alpha_1} + \frac{1}{\alpha_2} + \ldots + \frac{1}{\alpha_k} \leqslant 10$$

and, consequently,

$$\delta_a \leqslant \frac{0.5}{10^{n-2}}.$$

4

Thus, if all factors have n valid significant digits and the number of factors is not larger than 10, then the number of valid digits in the product is one or two units less than n. In the case when the factors have different accuracy, we must understand n to be the number of valid digits in the least exact factor.

Remark. When the number of factors is large ($k > 10$), it is convenient to use the statistical estimate which takes into account the partial compensation of the errors of unlike signs. Now if all the numbers a_i ($i = 1, 2, \ldots, k$) have approximately the same relative error δ, then the relative error of the product is assumed to be

$$\delta_a = \sqrt{n}\,\delta. \tag{11}$$

Example 3. Find the relative error and the number of valid digits of the product $a = 84.76 \cdot 8.436$, where all the digits of the factors are valid.

△ Here $a_1 = 84.76$, $a_2 = 8.436$, $n_1 = n_2 = 4$. The product $a = 715.03\ldots$ itself begins with the digit 7, i.e. its $\alpha_1 = 7$. Using now formula (10), we find that

$$\delta_a = \frac{0.5}{10^{4-1}} \cdot \left(\frac{1}{8} + \frac{1}{8}\right) = \frac{0.5}{4 \cdot 10^{4-1}}.$$

Comparing this result with the right-hand side of formula (4) from 1.5, we find that

$$\frac{0.5}{(7+1) \cdot 10^{4-1}} < \frac{0.5}{4 \cdot 10^{4-1}} < \frac{0.5}{(7+1) \cdot 10^{3-1}}.$$

Consequently, the product has at least three valid digits.

Let us verify whether this is so. We seek the absolute error using the formula $\Delta_a = |a|\,\delta_a$. We get $\Delta_a = 715.1 \cdot 0.125 \times 10^{-3} \cong 0.09$. It follows that the approximate value of the product has three valid digits and, with due account of the rounding error of the result, we can write

$$A = 715.0 \pm 0.2. \ \blacktriangle$$

Example 4. Find the relative error of the product $a = 145.35 \times 1.24386$ and the number of valid digits in it if the numbers are given with valid digits.

△ Here $a_1 = 145.35$, $n_1 = 5$, $a_2 = 1.24386$, $n_2 = 6$. These numbers have different numbers of valid significant digits. We choose $n = 5$. From formula (10) we obtain

$$\delta_a = \frac{0.5}{10^{5-1}} \left(\frac{1}{1} + \frac{1}{3}\right) = \frac{0.5}{7.5 \cdot 10^{4-1}}.$$

Having calculated the product $a = a_1 \cdot a_2 = 770.43\ldots$, we compare the quantity δ_a with the right-hand side of formula (4) from 1.5:

$$\frac{0.5}{(7+1) \cdot 10^{4-1}} < \frac{0.5}{7.5 \cdot 10^{4-1}} < \frac{0.5}{(7+1) \cdot 10^{3-1}}.$$

Consequently, the product has at least three valid significant digits. ▲

Thus, in an unfavourable case, the product of approximate numbers may have $n-2$ valid significant digits (where n is the least number of valid significant digits of the given factors).

1.8. The Error of a Quotient. The Number of Valid Digits of a Quotient

The error of a quotient. Let us consider exact numbers A_1, A_2 and their approximations a_1, a_2 with the absolute errors Δ_{a_1} and Δ_{a_2}. We pose the following problem: evaluate the relative error of the approximate value of the quotient $a = a_1/a_2$ for the exact value $A = A_1/A_2$. Let $a_1 \neq 0$, $a_2 \neq 0$. We represent the exact values of A_1 and A_2 in the form

$$A_1 = a_1 + \Delta_1, \quad A_2 = a_2 + \Delta_2, \tag{1}$$

where the unknowns Δ_1 and Δ_2 satisfy the inequalities

$$|\Delta_1| \leqslant \Delta_{a_1}, \quad |\Delta_2| \leqslant \Delta_{a_2}. \tag{2}$$

Let us consider now the difference

$$A - a = \frac{a_1 + \Delta_1}{a_2 + \Delta_2} \quad \frac{a_1}{a_2} = \frac{a_2\Delta_1 - a_1\Delta_2}{a_2(a_2+\Delta_2)}$$

Having divided the right-hand and left-hand sides by a, we consider their absolute values:

$$\frac{a_2\Delta_1 - a_1\Delta_2}{a_1(a_2+\Delta_2)} \qquad \frac{a_2}{a_2+\Delta_2} \quad \frac{\Delta_1}{a_1} \quad \frac{\Delta_2}{a_2}$$

Bearing in mind that Δ_2 is small as compared to a_2, we approximately set $a_2/(a_2+\Delta_2) \cong 1$. Then, using the properties of absolute values and inequalities (2), we obtain

$$\left|\frac{A-a}{a}\right| = \left|\frac{\Delta_1}{a_1} - \frac{\Delta_2}{a_2}\right| \leqslant \frac{\Delta_{a_1}}{|a_1|} + \frac{\Delta_{a_2}}{|a_2|} = \delta_{a_1} + \delta_{a_2}.$$

Thus we can take the sum of the relative errors of the dividend and the divisor

$$\delta_a = \delta_{a_1} + \delta_{a_2} \tag{3}$$

as the relative error of the quotient $a = a_1/a_2$.

When we use formula (3) to evaluate the relative error of the quotient, the most important contribution to that error is from the least exact (having the greatest relative error) numbers. Therefore, when the dividend and the divisor have different relative errors, we usually do the following:

(1) isolate the least exact number, i.e. the number with the least number of valid digits,

(2) round off the second number, leaving in it one significant digit more than there are digits in the isolated number,

(3) retain as many significant digits in the quotient as there were in the least exact number.

Knowing the relative error of the quotient, it is easy to find its absolute error using the formula

$$\Delta_a = |a|\ \delta_a = \left|\frac{a_1}{a_2}\right| (\delta_{a_1} + \delta_{a_2}). \tag{4}$$

Example 1. Calculate the quotient $a = x/y$ of the approximate numbers $x = 5.735$ and $y = 1.23$ if all the digits of the dividend and the divisor are valid. Find the relative and absolute errors.

△ (1) We shall first calculate the quotient. Since the dividend $x = 5.735$ has four valid significant digits and the divisor has three, we can carry out the division without rounding-off. We have $a = 5.735 \div 1.23 = 4.66$. We have retained three significant digits in the result since the least exact number (the divisor) contains three valid significant digits.

(2) Let us calculate the relative error of the quotient using formula (3) and bearing in mind that $\Delta_x = 0.0005$, $\Delta_y = 0.005$:

$$\delta_a = \delta_x + \delta_y = \frac{0.0005}{5.735} + \frac{0.005}{1.23} = 0.00009 + 0.0041 = 0.0042 \cong 0.5\%.$$

(3) We seek the absolute error

$$\Delta_a = |\ a\ |\ \delta_a = 4.66 \cdot 0.0042 = 0.02.$$

With due account of the rounding error 0.005, the final result must be written as $A = 4.66 \pm 0.03$. Note that the hundred's digit is doubtful since $0.03 > 0.005$. If we write the result only with valid significant digits, we must round it off and take into account the rounding error. This requirement is satisfied only by the approximate number $a_1 = 5$ since $\Delta_{a_1} = \Delta_a + \Delta^1_{\text{rounding}} = 0.02 + 0.4 = 0.42 < 0.5$.

At the same time, we cannot leave two valid significant digits in the approximate number a since then we would obtain $a_2 = 4.7$ and $\Delta_{a_2} = \Delta_a + \Delta^2_{\text{rounding}} = 0.02 + 0.05 = 0.07 > 0.05$ ▲

The number of valid digits in the quotient. Let the approximate numbers

$$a_1 = \alpha_1 \cdot 10^{l_1} + \alpha_2 \cdot 10^{l_1 - 1} + \ldots;$$

$$a_2 = \beta_1 \cdot 10^{l_2} + \beta_2 \cdot 10^{l_2 - 1} + \ldots$$

have n valid significant digits each. Then, using the equalities

$$\delta_{a_1} = \frac{0.5}{\alpha_1 \cdot 10^{n-1}}, \quad \delta_{a_2} = \frac{0.5}{\beta_1 \cdot 10^{n-1}},$$

we shall find the relative error of the quotient $a = a_1/a_2$.

$$\delta_a = \delta_{a_1} + \delta_{a_2} = \frac{0.5}{\alpha_1 \cdot 10^{n-1}} + \frac{0.5}{\beta_1 \cdot 10^{n-1}}$$

$$= \frac{0.5}{10^{n-1}} \left(\frac{1}{\alpha_1} + \frac{1}{\beta_1} \right). \tag{5}$$

Consequently, if $\alpha_1 \geqslant 2$ and $\beta_1 \geqslant 2$, then the quotient has at least $n - 1$ valid significant digits. Now if $\alpha_1 = 1$ or $\beta_1 = 1$, then the quotient may have $n - 2$ valid significant digits.

Example 2. Calculate the quotient $a = 39.356 \div 2.21$ and find out how many valid significant digits it contains if all the digits in the dividend and divisor are valid.

△ (1) Since the divisor has three valid significant digits and the dividend has five, we round off the dividend to four significant digits and make the division. We have $a = 39.36 \div 2.21 = 17.81 \cong 17.8$ (we leave as many significant digits in the result as there are in the number with the smaller number of valid significant digits).

(2) We can find the relative error using formula (5), where $n = 3$ since the least exact number has three valid digits; $\alpha_1 = 3$, $\beta_1 = 2$. Consequently,

$$\delta_a = \frac{0.5}{10^2} \left(\frac{1}{3} + \frac{1}{2} \right) = \frac{5}{12} \cdot 10^{-2} = 0.42\%.$$

Comparing the quantity δ_a with the right-hand side of formula (4) from 1.5, we obtain

$$\frac{0.5}{(1+1) \cdot 10^{3-1}} < \frac{0.5}{1.2 \cdot 10^{3-1}} < \frac{0.5}{(1+1) \cdot 10^{2-1}}.$$

Thus the quotient contains at least two valid significant digits, i.e. one significant digit less than the approximate number (the divisor) with the smaller number of valid significant digits.▲

Example 3. Find the relative error of the quotient $a = 15.834 \div 1.72$ and the number of valid digits in it if the dividend and divisor have valid significant digits.

△ The least exact number has three valid significant digits. We shall seek the relative error using formula (5):

$$\delta_a = \frac{0.5}{10^2}\left(\frac{1}{1}+\frac{1}{1}\right) : \frac{0.5\cdot 2}{10^{3-1}}$$

Calculating the quotient $a = 9.20$ and comparing the quantity δ_a with the right-hand side of formula (4) from 1.5, we find that the quotient contains only one valid significant digit, i.e. two valid significant digits less than the least exact number. ▲

1.9. The Errors of a Power and a Root

Consider an approximate number a_1 which has a relative error δ_{a_1}. Assume that we have to evaluate the relative error of degree $a = a_1^m$. Evidently,

$$a = a_1^m = \underset{m\ \text{factors}}{a_1 a_1 \ldots a_1}.$$

The relative error of the product is

$$\delta_a = \underbrace{\delta_{a_1} + \delta_{a_1} + \ldots + \delta_{a_1}}_{m\ \text{terms}} = m\delta_{a_1}. \qquad (1)$$

Thus, when we raise the approximate number a to the power m, its relative error increases m times. In practical calculations, when we raise an approximate number to a power, we leave as many significant digits in the result as there were in the approximate number itself.

Example 1. A side of the square $a = 36.5$ cm (with an accuracy to 1 mm). Find the area of the square, the relative and the absolute error and the number of valid digits in the result.

△ (1) We calculate the area of the square:

$$s = a^2 = 36.5^2 = 1332.2 \cong 1.33\cdot 10^3 \text{ cm}^2.$$

(2) We seek the relative error of the area:

$$\delta_s = 2\delta_a = 2\cdot\frac{0.1}{36.5} \cong 0.0055 = 0.55\%.$$

(3) Now we seek the absolute error of the area:

$$\Delta_s = s\delta_s = 1.33\cdot 10^3\cdot 0.0055 \cong 7.4 \text{ cm}^2.$$

With due account of the rounding error, we can write the final result as

$$s = (1.33 \pm 0.01)\cdot 010^3 \text{ m}^2.$$

Thus the result has two valid significant digits. ▲

Note that in the broad sense the result of Example 1 has three valid significant digits.

Let us consider an approximate number a_1 which has a relative error δ_{a_1}. We can show that the relative error of the number $a = \sqrt[m]{a_1}$ is m times as small as the relative error of the number a_1:

$$\delta_a = \frac{1}{m}\,\delta_{a_1}. \qquad (2)$$

In practical calculations, when we extract a root of an approximate number, we leave as many significant digits in the result as there were in the radicand.

Example 2. Find out with what relative error and with how many valid significant digits we can calculate a side of the square if its area $s = 16.45$ cm^2 with an accuracy of 0.01.

△ We have $a = \sqrt{s} = 4.056$ cm;

$$\delta_a = \frac{1}{2}\,\delta_s = \frac{1}{2}\cdot\frac{0.01}{16.45} = 0.0003 = 0.03\%;$$

$$\Delta_a = 4.056\cdot 0.0003 = 1.3\cdot 10^{-3}.$$

Thus, with due account of the rounding error, we have $A = 4.056 \pm 0.002$ cm and the number of valid significant digits is 3. ▲

1.10. The Rules of Calculating Digits

There are certain rules for calculating digits which should be used whenever the calculation of the error may not be strict. These rules show how all the results must be rounded off in order, first, to ensure the specified accuracy of the final result and, second, to obviate calculations with extra digits which do not affect the valid digits of the result.

Here are the **rules of calculating digits.**

1° When adding up and subtracting approximate numbers, we must leave as many decimal digits in the result as there are in the approximate number with the least number of decimal digits.

2°. When multiplying and dividing approximate numbers, we must retain as many significant digits in the result as there are in the approximate number with the least number of valid significant digits.

3°. When we square or cube an approximate number, we must retain as many significant digits in the result as there are in the base of the power.

4°. When we extract a square or cubic root of an approximate number, we must retain as many significant digits in the result as there are in the radicand.

5°. When calculating intermediate results, we must retain one digit more than rules 1°-4° recommend. In the final result this "reserve" digit is dropped.

6°. If some data have more decimal digits (in addition and subtraction) or more significant digits (in other operations) than other data, they must be first rounded off, with only one "reserve" digit retained.

7°. When we use logarithms to calculate a one-term expression, we should calculate the number of significant digits in the given approximate number which has the least number of significant digits and use the table of logarithms which has one extra decimal digit. In the final result we drop the last significant digit.

8°. If we can take the data with an arbitrary accuracy, then, to obtain a result with m valid digits, we must take the initial data with the number of digits which, according to the preceding rules, ensure $m+1$ digit in the result.

We present these rules on the assumption that the components of the operations contain only valid digits and the number of operations is not large.

Example 1. Calculate $X=\frac{A^3\sqrt{B}}{C^2}$, where $A=7.45\pm 0.01$, $B=50.46\pm 0.02$, $C=15.4\pm 0.03$. Find the error of the result.

△ When we calculate the intermediate results, we shall retain one "reserve" digit, i.e. if, according to the general rule we must retain n significant digits, we shall retain $n+1$ digits in the intermediate results. We have

$$a^3=413.5,\quad \sqrt{b}=7.1035,\quad c^2=237.2,\quad x=\frac{413.5\cdot 7.1035}{237.2}=12.4.$$

We have left three significant digits in the result since the least number of significant digits in the factors is 3.

We calculate the error of the result:

$$\delta_x=3\delta_a+\frac{1}{2}\,\delta_b+2\delta_c=3\cdot\frac{0.01}{7.45}+\frac{1}{2}\cdot\frac{0.02}{50.46}+2\cdot\frac{0.03}{15.4}\cong 0.0041+0.0002+0.004\cong 0.009;$$

$$\Delta_x = 12.4 \cdot 0.009 \cong 0.12.$$

The answer is $X = 12.4 \pm 0.12$, $\delta_x = 0.9\%$. ▲

Example 2. Calculate $X = \frac{(A+B)M}{(C-D)^2}$, where $A = 2.754 \pm 0.001$, $B = 11.7 \pm 0.04$, $M = 0.56 \pm 0.05$, $C = 10.536 \pm 0.002$, $D = 6.32 \pm 0.008$. Find the errors of the result.

△ We find that

$$a + b = 2.75 + 11.7 = 14.45,$$

$\Delta_{a+b} = \Delta_a + \Delta_b + \Delta_{\text{rounding}} = 0.001 + 0.04 + 0.004 = 0.045$, $c - d = 10.536 - 6.32 = 4.216$; $\Delta_{c-d} = 0.002 + 0.008 = 0.010$. Therefore

$$x = \frac{14.45 \cdot 0.56}{4.216^2} = \frac{14.45 \cdot 0.56}{17.75} = 0.456 \cong 0.46 = 4.6 \cdot 10^{-1};$$

$$\delta_x = \frac{0.045}{14.45} + \frac{0.005}{0.56} + 2 \cdot \frac{0.01}{4.216}$$

$$= 0.00311 + 0.00894 + 0.00474 = 0.02 = 2\%.$$

Consequently,

$$\Delta_x = 0.46 \cdot 0.02 = 0.01.$$

Thus the result is $X = 0.46 \pm 0.01$, $\delta_x = 2\%$. ▲

Example 3. Using the rules of calculating digits, compute

$$v = \pi h^2 \left(r - \frac{h}{3}\right),$$

where $h = 11.8$, $\pi = 3.142$, $r = 23.67$.

△ We find that

$$v = 3.142 \cdot 11{,}8^2\,(23.67 - 3.933) = 3.142 \cdot 11.8^2 \cdot 19.737$$
$$= 3.142 \cdot 139.2 \cdot 19.737 = 437.37 \cdot 19.737 = 8630 \cong 8.63 \cdot 10^3.$$ ▲

Exercises

1. Carry out the successive roundings-off of the following numbers: (a) 2.75464, (b) 3.14159, (c) 0.56453, (d) 4.1945, (e) 0.60653.

2. Rounding off the following numbers to three significant digits, find the absolute Δ_a and the relative (in per cent) δ_a error of the resulting approximations: (a) 1.1426, (b) 0.01015, (c) 0.1245, (d) 921.55, (e) 0.002462.

3. Find the absolute error Δ_x of the following approximate numbers proceeding from their relative error: (a) $x = 2.52$, $\delta_x = 0.7\%$, (b) $x = 0.986$, $\delta_x = 10\%$, (c) $x = 46.75$, $\delta_x = 1\%$, (d) $x = 199.1$, $\delta_x = 0.01$, (e) $x = 0.86341$, $\delta_x = 0.0004$.

4. Find the number of valid significant digits for the following approximate numbers: (a) 39.285 ± 0.034, (b) 1.2785 ± 0.0007, (c) 183.3 ± 0.1, (d) 0.056 ± 0.0003, (e) 84.17 ± 0.0073.

5. Find out which of the following two equalities is more exact: (a) $6/25 \cong 1.4$ or $1/3 \cong 0.333$, (b) $1/9 \cong 0.1$ or $1/3 \cong 0.33$, (c) $15/7 \cong 2.14$ or $1/9 \cong 0.11$, (d) $6/7 \cong 0.86$ or $\pi \cong 22/7$, (e) $\pi \cong 3.142$ or $\sqrt{10} \cong 3.1623$.

Hint. First find the relative errors. The equality whose relative error is smaller is more exact.

6. Round off the doubtful digits of the number $A = 47.453 \pm 0.024$ leaving valid digits in it.

7. Round off the doubtful digits in the number $A = 46.3852 \pm 0.0031$ leaving valid digits in it.

8. Round off the doubtful digits in the approximate number $a = 3.2873$ if $\delta_a = 0.1\%$, leaving valid digits in it.

9. Find the relative and the absolute errors of the following approximate numbers if they have only valid digits: (a) $a = 0.7538$, (b) $a = 17.354$.

Hint. Use formula (3) from 1.5.

10. Calculate the following expressions and evaluate their errors. In the answer retain all valid digits and one doubtful digit. All the numbers are given with valid digits:

$$\text{(a)}\quad y=\frac{3.07\cdot 326}{36.4\cdot 323},\qquad \text{(b)}\quad y=\frac{36\cdot 245\cdot 85}{975\cdot 642},$$

$$\text{(c)}\quad y=\frac{37.2+458.67}{36.5\cdot 246},\qquad \text{(d)}\quad y=\frac{96.891-4.25}{33.3+0.426}.$$

11. Using the rules of calculating digits, compute:

$$\text{(a)}\quad s=\frac{L^2}{18}\cdot\frac{a^2+4ab+b^2}{(a+b)^2},\ \text{where}\ h=2.73,\quad a=0.152,\quad b=0.328;$$

$$\text{(b)}\quad s=\frac{1}{4}\pi\,(a^2-b^2),\ \text{where}\ a=0.937,\quad b=0.0630.$$

Chapter 2

Matrix Algebra and Some Data from the Theory of Linear Vector Spaces

2.1. Matrices and Vectors. Principal Operations Involving Matrices and Vectors

A rectangular array which is composed of elements (numbers in special cases) and has m rows and n columns is a *matrix* of dimension $m \times n$, or an $m \times n$ matrix. The elements of a matrix are designated as a_{ij}, where i is the number of the row and j is the number of the column whose intersection is occupied by the element.

For instance,

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix}$$

is an $m \times n$ matrix which has m rows and n columns.

The abridged notation of the matrix is $A = [a_{ij}]$, where $i = 1, 2, \dots, m$, $j = 1, 2, \dots, n$, or $[a_{ij}]_{mn}$.

If the number of rows in a matrix is not equal to the number of columns, i.e. $m \neq n$, then the matrix is called *rectangular*.

A matrix which has only one row, i.e. $m = 1$ is a *row matrix* (or *row vector*), for example,

$$A = [a_{11}a_{12} \qquad\qquad _{\imath}] \text{ or } A = [1\ \ 2\ \ 3\ \ 4].$$

A matrix which has only one column, i.e. $n = 1$, is a *column matrix* (or *column vector*), for example,

$$A = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} \quad \text{or} \quad A = \begin{bmatrix} 1 \\ 0 \\ 4 \\ 5 \end{bmatrix}.$$

In what follows, we shall call a row matrix or a column matrix a *vector* and designate it as $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$ or as $\mathbf{x} = [x_1, x_2, \ldots, x_n]$. The numbers $x_1, x_2, \ldots, x_n$ are the *coordinates* (or *elements*) of the vector $\mathbf{x}$. Since the number of coordinates of a vector is, by definition, its dimension, the vector $\mathbf{x}$ is n-dimensional.

If the number of rows in a matrix is equal to the number of columns, i.e. $m = n$, the matrix is *square*. A matrix of this kind can be written in the form

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}.$$

For a square matrix the total number of rows or columns is called the *order* of the matrix.

The *principal diagonal* of a square matrix is the diagonal extending from the upper left to the lower right corner, i.e. a set of elements of the form a_{ii}, where $i = 1, 2, \ldots, n$.

A square matrix in which all the elements which lie outside of the principal diagonal are zero, is called a *diagonal matrix*. This matrix has the form

$$A = \begin{bmatrix} a_{11} & 0 & 0 \ldots 0 \\ 0 & a_{22} & 0 \ldots 0 \\ 0 & 0 & a_{33} \ldots 0 \\ \cdot & \cdot & \cdot \\ 0 & 0 & 0 \ldots a_{nn} \end{bmatrix}.$$

A diagonal matrix all of whose elements on the principal diagonal are equal to unity is an *identity* (or *unit*) *matrix*. An identity matrix is designated as I and has the form

$$I_n = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix}.$$

The index n indicates the order of the identity matrix. In matrix calculation the identity matrix plays the part of unity.

A square matrix all of whose elements are symmetric about the principal diagonal is a *symmetric matrix*. The equality $a_{ij} = a_{ji}$ $(i \neq j)$ holds true for a symmetric matrix. For example,

$$A = \begin{bmatrix} 1 & 2 & 4 \\ 2 & 3 & 5 \\ 4 & 5 & 6 \end{bmatrix}$$

is a symmetric matrix.

A matrix all of whose elements are zero is a *zero*, or *null*, matrix and is designated as 0. If we have to indicate the number of rows and columns, we write

$$0_{mn} = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & 0 \end{bmatrix}.$$

Two matrices $A = [a_{ij}]$ and $B = [b_{ij}]$ are *equal* if (1) they are of the same dimension, i.e. the number of rows of the matrix A is equal to the number of rows of the matrix B and the number of columns of the matrix A is equal to the number of columns of the matrix B, (2) the respective elements of these matrices are equal. Thus, if

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}, \quad B = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ b_{m1} & b_{m2} & \dots & b_{mn} \end{bmatrix}$$

and $a_{ij} = b_{ij}$ $(i = 1, 2, \dots, m,\ j = 1, 2, \dots, n)$, then $A = B$.

The *sum* of two matrices of the same dimension $A + B = [a_{ij}] + [b_{ij}]$ $(i = 1, 2, \dots, m,\ j = 1, 2, \dots, n)$ is a matrix $C = [c_{ij}]$ of the same dimension whose elements c_{ij} are equal to the sums of the respective elements a_{ij} and b_{ij} of the matrices A and B, i.e. $c_{ij} = a_{ij} + b_{ij}$.

The *difference* of matrices is defined by analogy with the sum of matrices with the only difference that the signs of the elements of the subtracted matrix are changed to the opposite, i.e. $D = A - B$, $d_{ij} = a_{ij} - b_{ij}$ ($i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$).

The *product of the matrix* $A = [a_{ij}]$ *by the number* α is a matrix whose elements result from the multiplication of all the elements of the matrix A by the number α:

$$\alpha A = \begin{bmatrix} \alpha a_{11} & \alpha a_{12} & \ldots & \alpha a_{1n} \\ \alpha a_{21} & \alpha a_{22} & \ldots & \alpha a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \alpha a_{m1} & \alpha a_{m2} & \ldots & \alpha a_{mn} \end{bmatrix}.$$

The matrix $-A = (-1)A$ is the *inverse* of the matrix A. The addition of matrices obeys the following laws: (1°) $A + (B + C) = (A + B) + C$, (2°) $A + B = B + A$, (3°) $A + 0 = A$, (4°) $A + (-A) = 0$.

The product of a matrix by a number obeys the following laws:

(1°) $1 \cdot A = A$, (2°) $0 \cdot A = 0$, (3°) $\alpha(\beta A) = (\alpha\beta)A$.

Example 1. Assume that

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ -2 & 0 & 4 & 5 \\ -7 & 6 & 1 & 2 \\ 1 & -1 & 4 & 3 \end{bmatrix}; \quad B = \begin{bmatrix} 1 & -1 & 2 & -7 \\ 4 & 3 & -2 & -4 \\ 1 & -6 & 0 & -1 \\ 2 & 3 & 4 & -5 \end{bmatrix}.$$

Then

$$C = A + B = \begin{bmatrix} 2 & 1 & 5 & -3 \\ 2 & 3 & 2 & 1 \\ -6 & 0 & 1 & 1 \\ 3 & 2 & 8 & -2 \end{bmatrix}; \quad D = A - B = \begin{bmatrix} 0 & 3 & 1 & 11 \\ -6 & -3 & 6 & 9 \\ -8 & 12 & 1 & 3 \\ -1 & -4 & 0 & 8 \end{bmatrix}$$

Example 2. The product of the matrix

$$A = \begin{bmatrix} 1 & 3 & -4 \\ 0 & 1 & -2 \\ -3 & 1 & 5 \end{bmatrix}$$

by the number 2 is a matrix

$$2A = \begin{bmatrix} 2 & 6 & -8 \\ 0 & 2 & -4 \\ -6 & 2 & 10 \end{bmatrix}.$$

The *product AB of two matrices*

$$A=\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \text{ and } B=\begin{bmatrix} b_{11} & b_{12} & \dots & b_{1q} \\ b_{21} & b_{22} & \dots & b_{2q} \\ \cdot & \cdot & \cdot & \cdot \\ b_{n1} & b_{n2} & \dots & b_{nq} \end{bmatrix},$$

which have dimensions $m \times n$ and $n \times q$ respectively, is a matrix

$$C=\begin{bmatrix} c_{11} & c_{12} & \dots & c_{1q} \\ c_{21} & c_{22} & \dots & c_{2q} \\ \cdot & \cdot & \cdot & \cdot \\ c_{m1} & c_{m2} & \dots & c_{mq} \end{bmatrix}$$

of dimension $m \times q$. Note that the matrix $C = AB$ is defined only when the number of columns of the matrix A is equal to the number of rows of the matrix B.

The elements of the matrix C can be calculated according to the following rule: *to obtain the element c_{ij} which is in the ith row and the jth column of the product of two matrices, the elements of the ith row of the first matrix must be multiplied by the corresponding elements of the jth column of the second matrix and the resulting products summed up*:

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \dots + a_{in}b_{nj}$$

$$(i = 1, 2, \dots, m,\ j = 1, 2, \dots, q).$$

For example, $c_{23} = a_{21}b_{13} + a_{22}b_{23} + \dots + a_{2n}b_{n3}$, $c_{41} = a_{41}b_{11} + a_{42}b_{21} + \dots + a_{4n}b_{n1}$ etc.

Example 3. $AB=\begin{bmatrix} 3 & 2 & 8 & 1 \\ 1 & -4 & 0 & 3 \end{bmatrix}\begin{bmatrix} 2 & -1 \\ 1 & -3 \\ 0 & 1 \\ 3 & 1 \end{bmatrix}$

$$=\begin{bmatrix} 3\cdot2+2\cdot1+8\cdot0+1\cdot3 & 3(-1)+2\cdot(-3)+8\cdot1+1\cdot1 \\ 1\cdot2+(-4)\cdot1+0\cdot0+3\cdot3 & 1(-1)+(-4)\cdot(-3)+0\cdot1+3\cdot1 \end{bmatrix}$$

$$=\begin{bmatrix} 11 & 0 \\ 7 & 14 \end{bmatrix}.$$

Here $AB = [a_{ij}]_{2\times4}\cdot[b_{ij}]_{4\times2} = C = [c_{ij}]_{2\times2}$.

Example 4.

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 1\cdot1+2\cdot2+3\cdot3 \\ 4\cdot1+5\cdot2+6\cdot3 \\ 7\cdot1+8\cdot2+9\cdot3 \end{bmatrix} = \begin{bmatrix} 14 \\ 32 \\ 50 \end{bmatrix}.$$

Here $AB = [a_{ij}]_{3\times3}\,[b_{ij}]_{3\times1} = [c_{ij}]_{3\times1}$.

Example 5.

$$AB = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 1\cdot5+2\cdot7 & 1\cdot6+2\cdot8 \\ 3\cdot5+4\cdot7 & 3\cdot6+4\cdot8 \end{bmatrix} = \begin{bmatrix} 19 & 22 \\ 43 & 50 \end{bmatrix},$$

$$BA = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 5\cdot1+6\cdot3 & 5\cdot2+6\cdot4 \\ 7\cdot1+8\cdot3 & 7\cdot2+8\cdot4 \end{bmatrix} = \begin{bmatrix} 23 & 34 \\ 31 & 46 \end{bmatrix},$$

i.e. $AB \neq BA$.

Example 6. The product $AB = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 3 & 2 & 4 \\ 2 & 1 & 3 \\ 4 & 3 & 0 \end{bmatrix}$

$$= \begin{bmatrix} 1\cdot3+2\cdot2+3\cdot4 & 1\cdot2+2\cdot1+3\cdot3 & 1\cdot4+2\cdot3+3\cdot0 \\ 4\cdot3+5\cdot2+6\cdot4 & 4\cdot2+5\cdot1+6\cdot3 & 4\cdot4+5\cdot3+6\cdot0 \end{bmatrix}$$

$$= \begin{bmatrix} 19 & 13 & 10 \\ 46 & 31 & 31 \end{bmatrix},$$

$$BA = \begin{bmatrix} 3 & 2 & 4 \\ 2 & 1 & 3 \\ 4 & 3 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

does not exist since the number of columns of the matrix A is not equal to the number of rows of the matrix B.

The product of matrices obeys the following laws:

(1°) $A\,(BC) = (AB)\,C$, (2°) $\alpha\,(AB) = (\alpha A)\,B$,

(3°) $(A + B)\,C = AC + BC$, (4°) $EA = A$.

Note that $AB \neq BA$, i.e. in the general case the product of two matrices does not possess the property of commutativity. The only exception is an identity matrix. $AI = IA = A$. This means that in the general case we cannot interchange the factors without changing the product. When we change the order of the factors, it may turn out that it is impossible to carry out the multiplication at all (see Example 6).

We say of the product AB of two matrices A and B that the matrix B is premultiplied by the matrix A and the matrix A is postmultiplied by the matrix B.

The product of several matrices ABC must be understood as follows: the matrix A is postmultiplied by the matrix B and the resulting matrix is postmultiplied by the matrix C, etc. The number of multiplied matrices may be arbitrary, the only condition being that two adjacent matrices can be multiplied.

The matrix A^n is the *nth power* of the matrix A. If A is a square matrix and n is a positive integer, then

$$A^n = \underbrace{A \cdot A \cdot A \ \ldots\ A}_{n \text{ factors}}.$$

The operations of addition of column matrices and row matrices (i.e. vectors) and multiplication of them by scalars are similar to the corresponding operations involving square matrices. Thus the *sum of two vectors* $\mathbf{x} = [x_1 x_2 \ldots x_n]$ and $\mathbf{y} = [y_1 y_2 \ldots y_n]$ is a vector $\mathbf{z} = [z_1 z_2 \ldots z_n]$ with the coordinates $z_1 = x_1 + y_1$, $z_2 = x_2 + y_2, \ldots, z_n = x_n + y_n$, the *product of the vector* $\mathbf{x} = [x_1 x_2 \ldots x_n]$ *by the scalar* α is a vector $\mathbf{z} = \alpha x = [\alpha x_1\ \alpha x_2 \ldots \alpha x_n]$.

Example 7. The sum of the vectors $\mathbf{x} = [1\ 2\ 3]$ and $\mathbf{y} = [-5\ -2\ 4]$ is a vector $\mathbf{z} = [-4\ 0\ 7]$; the product of the vector $\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ -3 \end{bmatrix}$ by the scalar $\alpha = 2$ is a vector $\mathbf{z} = \begin{bmatrix} 2 \\ 4 \\ -6 \end{bmatrix}$.

2.2. Transpose of a Matrix

If we replace the rows of the $m \times n$ matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot\ \cdot\ \cdot & \cdot \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix}$$

by the corresponding columns, we obtain an $n \times m$ matrix

$$A^{\mathrm{T}} = \begin{bmatrix} a_{11} & a_{21} & \ldots & a_{m1} \\ a_{12} & a_{22} & \ldots & a_{m2} \\ \cdot & \cdot & \cdot\ \cdot\ \cdot & \cdot \\ a_{1n} & a_{2n} & \ldots & a_{mn} \end{bmatrix},$$

which is known as the *transpose* of the matrix A.

Example 1. The transpose of the 3×4 matrix

$$A=\begin{bmatrix}1 & 2 & -3 & 5\\ 2 & 0 & 4 & 2\\ 1 & 2 & 3 & 4\end{bmatrix}$$

is a matrix

$$A^{\mathrm{T}}=\begin{bmatrix}1 & 2 & 1\\ 2 & 0 & 2\\ -3 & 4 & 3\\ 5 & 2 & 4\end{bmatrix}$$

of dimension 4×3.

Example 2. The transpose of the row matrix $B=[1\ 2\ 3\ 4]$ is a column matrix $B^{\mathrm{T}}=\begin{bmatrix}1\\ 2\\ 3\\ 4\end{bmatrix}$.

Note the following properties of transposition.

1°. If we transpose the matrix A twice, it will remain unchanged:

$$(A^{\mathrm{T}})^{\mathrm{T}}=A.$$

2°. The transpose of the sum of two matrices is equal to the sum of the transposed matrices:

$$(A+B)^{\mathrm{T}}=A^{\mathrm{T}}+B^{\mathrm{T}}.$$

This follows from the definition of the sum of two matrices.

3°. The transpose of the product of two matrices is equal to the product of the transposed matrices taken in the reverse order:

$$(AB)^{\mathrm{T}}=B^{\mathrm{T}}A^{\mathrm{T}}.$$

The matrix $(AB)^{\mathrm{T}}$ has resulted from the multiplication of the elements of the rows of the matrix A by the elements of the columns of the matrix B followed by the replacement of the rows by the columns. We can get the same result if we multiply the elements of the columns of the matrix B (the rows of B^{T}) by the elements of the rows of the matrix A (the columns of A^{T}).

2.3. The Determinant of a Matrix. The Properties of the Determinant and the Rules of Its Calculation

Let A be an arbitrary square matrix of order n:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}.$$

The matrix A is associated with a *determinant* which can be designated as d, D, det A or $|A|$:

$$d = D = \det A = |A| = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix}. \tag{1}$$

The determinant of a matrix is a number which can be calculated in accordance with certain rules considered below.

There are two diagonals in a determinant, a principal, or leading, diagonal and a secondary diagonal. The *principal diagonal* of a determinant, as that of a square matrix, consists of the elements a_{ii}, where $i = 1, 2, \dots, n$. The *secondary diagonal* is perpendicular to the principal diagonal and passes from the upper right corner of the determinant to its lower left corner. The order of a determinant corresponds to the order of the matrix with which it is associated.

If the order of a matrix is unity, i.e. the matrix consists of one element a_{11}, then the *first-order determinant* associated with this matrix is a number equal to that element.

Consider a second-order square matrix

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

The *second-order determinant* corresponding to this matrix is a number

$$\det A = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}. \tag{2}$$

Formula (2) gives the rule of calculating a second-order determinant: *a second-order determinant is equal to the product of the elements of the principal diagonal minus the product of the elements of the secondary diagonal.*

Example 1. Calculate the determinant of the matrix $A = \begin{bmatrix} 1 & 2 \\ 4 & 5 \end{bmatrix}$.

$$\triangle \quad \det A = \begin{vmatrix} 1 & 2 \\ 4 & 5 \end{vmatrix} = 1 \cdot 5 - 4 \cdot 2 = -3. \quad \blacktriangle$$

A *third-order determinant* is a number

$$\det A = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$
$$= a_{11}a_{22}a_{33} - a_{13}a_{21}a_{32} + a_{12}a_{23}a_{31}$$
$$- a_{13}a_{22}a_{31} - a_{21}a_{12}a_{33} - a_{32}a_{33}a_{11}. \quad (3)$$

Thus *every term of a third-order determinant is the product of three of its elements, taken one from each row and each column. These products are taken with definite signs: three terms consisting of elements of the principal diagonal and of the element which are at the vertices of isosceles triangles with the bases parallel to the principal diagonal are taken with the plus sign* (Fig. 2.1); *three terms which occupy similar positions relative to the secondary diagonal are taken with the minus sign* (Fig. 2.2). This rule is known as the **rule of triangles**.

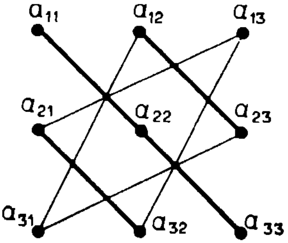


Fig. 2.1

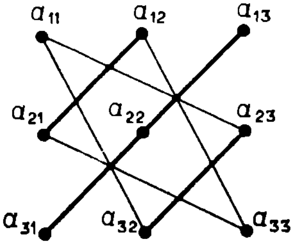


Fig. 2.2

Example 2. Calculate the determinant of the matrix $A = \begin{bmatrix} 1 & 2 & 3 \\ -4 & 5 & -1 \\ 2 & 1 & 2 \end{bmatrix}$.

$$\triangle \ \det A = \begin{vmatrix} 1 & 2 & 3 \\ -4 & 5 & -1 \\ 2 & 1 & 2 \end{vmatrix} = 1\cdot5\cdot2 + 3\cdot1\cdot(-4) + 2\cdot(-1)\cdot2$$

$$-2\cdot5\cdot3 - 1\cdot(-1)\cdot1 - 2\cdot(-4)\cdot2 = -19. \ \blacktriangle$$

Let us consider now a determinant of any order n, where $n \geqslant 2$. To calculate such a determinant, we must introduce the concepts of a minor and a cofactor.

The *minor of the element* a_{ij} of a determinant of order n (1) is a determinant of order $(n - 1)$ which can be obtained from the initial determinant by deleting the ith row and the jth column (the row and the column whose intersection is occupied by the element a_{ij}).

The minor of the element a_{ij} is designated as M_{ij}. The first index here denotes the number of the row and the second, the number of the column which are deleted.

For instance, in the third-order determinant

$$d = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

the second-order determinant

$$M_{12} = \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix}$$

is the minor of the element a_{12}.

The *cofactor*, or *algebraic adjunct*, of the element a_{ij} of the nth-order determinant (1) is a number

$$A_{ij} (-1)^{i+j} M_{ij}.$$

Evidently, if the sum $i + j$ is even, then the cofactor has the same sign as the minor, now if the sum $i + j$ is odd, then the sign is changed to the opposite.

Theorem 1. *A determinant is equal to the sum of the products of the elements of any row (column) by the corresponding cofactor:*

$$\det A = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix}$$

$$= a_{i1}A_{i1} + a_{i2}A_{i2} + \dots + a_{in}A_{in} = \sum_{j=1}^{n} a_{ij}A_{ij}$$

$$(i = 1, 2, \dots, n) \quad (4)$$

or

$$\det A = a_{1j}A_{1j} + a_{2j}A_{2j} + \dots + a_{nj}A_{nj}$$

$$= \sum_{i=1}^{n} a_{ij}A_{ij} \ (j = 1, 2, \dots, n). \quad (5)$$

Formula (4) is the *expansion of the determinant according to the elements of the ith row* and formula (5) is the *expansion of the determinant according to the elements of the jth column*.

When we expand a second-order determinant according to the elements of any row (column), we get formula (2) given above, and when we expand a third-order determinant according to the elements of any row (column), we get formula (3) (the rule of triangles).

Example 3. Calculate the determinant $d = \begin{vmatrix} 1 & 2 \\ 3 & 4 \end{vmatrix}$ by expanding it according to the elements of the first row.

△ According to formula (4) we have

$$d = a_{11}A_{11} + a_{12}A_{12}.$$

Since $A_{11} = (-1)^{1+1} \cdot 4 = 4$, $A_{12} = (-1)^{1+2} \cdot 3 = -3$, we hav $d = 1 \cdot 4 + 2(-3) = -2$. ▲

Example 4. Calculate the determinant $d = \begin{vmatrix} 3 & 2 & 1 \\ 2 & 5 & 3 \\ 3 & -1 & 2 \end{vmatrix}$, expanding it according to the elements of the second column.

△ From formula (5) we get $d = a_{12}A_{12} + a_{22}A_{22} + a_{32}A_{32}$. Next we find that

$$A_{12} = (-1)^{1+2} \begin{vmatrix} 2 & 3 \\ 3 & 2 \end{vmatrix} = -(4-9) = 5;$$

$$A_{22} = (-1)^{2+2} \begin{vmatrix} 3 & 1 \\ 3 & 2 \end{vmatrix} = 6 - 3 = 3;$$

$$A_{32} = (-1)^{3+2} \begin{vmatrix} 3 & 1 \\ 2 & 3 \end{vmatrix} = -(9-2) = -7;$$

whence $d = 2 \cdot 5 + 5 \cdot 3 + 4(-7) = -3$. ▲

Theorem 2 (corollary of Theorem 1). *If all of the elements of the ith row (column) of the determinant d, except for one, say, a_{ik}, are zero, then the determinant is equal to the product of the element a_{ik} by its cofactor,* i.e.

$$d = a_{ik}A_{ik}. \tag{6}$$

Example 5. Calculate the fourth-order determinant

$$d = \begin{vmatrix} 1 & 1 & 3 & 4 \\ 1 & 0 & 0 & 4 \\ 3 & 0 & 0 & 2 \\ 0 & 0 & -5 & -11 \end{vmatrix},$$

expanding it according to the elements of the second column.

△ Since $a_{22} = a_{32} = a_{42} = 0$, we find from formula (6) that

$$d = a_{12}A_{12} = 1\cdot(-1)^{1+2}\begin{vmatrix} 1 & 0 & 4 \\ 3 & 0 & 2 \\ 0 & -5 & -11 \end{vmatrix},$$

whence, again expanding the resulting third-order determinant according to the elements of the second column, we obtain

$$d = -(-5)(-1)^{3+2}\begin{vmatrix} 1 & 4 \\ 3 & 2 \end{vmatrix} = -5(2-12) = 50. \ ▲$$

Theorem 3. *The sum of the products of the elements of a row or a column of a determinant by the cofactors of the corresponding elements of the parallel row (or column) is zero.*

Thus, for the third-order determinant

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

there hold equalities $a_{21}A_{11} + a_{22}A_{12} + a_{23}A_{13} = 0$, $a_{31}A_{21} + a_{32}A_{22} + a_{33}A_{23} = 0$ and so on.

Here are the properties of the determinant.

1°. *The determinant does not change upon transposition:*

$$\det A = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix} = \begin{vmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ \cdot & \cdot & \cdot & \cdot \\ a_{1n} & a_{2n} & \dots & a_{nn} \end{vmatrix}.$$

This means that the rows and the columns of a determinant are equivalent.

It follows from this property that the *determinant of the matrix A is equal to the determinant of the transposed matrix A^{T}.*

For example,

$$\det A=\begin{vmatrix}1 & 2 & 3\\ -4 & 5 & -1\\ 2 & 1 & 2\end{vmatrix}=\begin{vmatrix}1 & -4 & 2\\ 2 & 5 & 1\\ 3 & 1 & 2\end{vmatrix}$$

$$=1\cdot5\cdot2+3\cdot(-4)\cdot1+2\cdot(-1)\cdot2-3\cdot5\cdot2-1\cdot(-1)\cdot1$$
$$-2\cdot(-4)\cdot2=-19.$$

2°. *If one of the rows or one of the columns of a determinant consists of zeros, then the determinant is equal to zero.*

For example,

$$\det A=\begin{vmatrix}0 & 1 & 2\\ 0 & 3 & 4\\ 0 & 10 & 15\end{vmatrix}$$

$$=0\cdot3\cdot15+0\cdot1\cdot4+0\cdot2\cdot10-0\cdot3\cdot2$$
$$-0\cdot1\cdot15-0\cdot4\cdot10=0.$$

3°. *When two rows or two columns are interchanged, the determinant only changes sign.*

For example

$$\det A=\begin{vmatrix}1 & 2 & 3\\ 2 & 1 & 2\\ -4 & 5 & -1\end{vmatrix}$$

$$=1\cdot1\cdot(-1)+2\cdot2\cdot(-4)+2\cdot5\cdot3-3\cdot1\cdot(-4)$$
$$-5\cdot2\cdot1-2\cdot2\cdot(-1)=19$$

(compare with the example which illustrated property 1°).

4°. *A determinant which contains two identical rows or two identical columns is equal to zero.*

For example,

$$\det A=\begin{vmatrix}1 & 2 & 3\\ 1 & 2 & 3\\ 2 & 1 & 2\end{vmatrix}=1\cdot2\cdot2+2\cdot3\cdot2+1\cdot1\cdot3-2\cdot2\cdot3-1\cdot2\cdot2-1\cdot1\cdot3=0.$$

5°. *If all the elements of a row or a column of a determinant are multiplied by the scalar $k\neq0$, then the determinant itself will be multiplied by that scalar.*

Here is another formulation of this property: *the common multiple of all the elements of some row or some column can be taken outside of the sign of the determinant.*

For example,

$$\det A=\begin{vmatrix}3&6&9\\2&1&2\\-4&5&-1\end{vmatrix}=3\begin{vmatrix}1&2&3\\2&1&2\\-4&5&-1\end{vmatrix}=3\cdot19=57$$

(compare with the example which illustrates property 3°).

6°. *A determinant which contains two proportional rows is equal to zero.*

For example,

$$\det A=\begin{vmatrix}3&6&9\\1&2&3\\2&1&2\end{vmatrix}=3\begin{vmatrix}1&2&3\\1&2&3\\2&1&2\end{vmatrix}=3\cdot0=0.$$

7°. *If all the elements of the ith row of an nth-order determinant are represented as the sum of two terms, i.e. $a_{ij}=b_{ij}+c_{ij}$ $(j=1, 2, \ldots, n)$, then the determinant is equal to the sum of two determinants in which all the rows, except for the ith row, are the same as those of the given determinant, and the ith row in one of the terms consists of the elements b_{ij} and in the other terms, of the elements c_{ij}:*

$$\det A$$

$$=\begin{vmatrix}a_{11}&a_{12}&a_{13}&\ldots&a_{1n}\\a_{21}=b_{21}+c_{21}&a_{22}=b_{22}+c_{22}&a_{23}=b_{23}+c_{23}&\ldots&a_{2n}=b_{2n}+c_{2n}\\a_{31}&a_{32}&a_{33}&\ldots&a_{3n}\\\cdot&\cdot&\cdot&\cdot&\cdot\\a_{n1}&a_{n2}&a_{n3}&\ldots&a_{nn}\end{vmatrix}$$

$$=\begin{vmatrix}a_{11}&a_{12}&a_{13}&\ldots&a_{1n}\\b_{21}&b_{22}&b_{23}&\ldots&b_{2n}\\a_{31}&a_{32}&a_{33}&\ldots&b_{3n}\\\cdot&\cdot&\cdot&\cdot&\cdot\\a_{n1}&a_{n2}&a_{n3}&\ldots&a_{nn}\end{vmatrix}+\begin{vmatrix}a_{11}&a_{12}&a_{13}&\ldots&a_{1n}\\c_{21}&c_{22}&c_{23}&\ldots&c_{2n}\\a_{31}&a_{32}&a_{33}&\ldots&a_{3n}\\\cdot&\cdot&\cdot&\cdot&\cdot\\a_{n1}&a_{n2}&a_{n3}&\ldots&a_{nn}\end{vmatrix}.$$

For example,

$$\det A=\begin{vmatrix}2&7&9\\1&3&4\\1&0&2\end{vmatrix}=\begin{vmatrix}1&3&4\\1&3&4\\1&0&2\end{vmatrix}+\begin{vmatrix}1&4&5\\1&3&4\\1&0&2\end{vmatrix}=0-1=-1.$$

8°. *If one of the rows of a determinant is the sum of some other rows or the sum of the products of some other rows of the determinant by the scalar k, then the determinant is* *equal to zero.* (This follows from properties 6° and 7° of a determinant.)

For example,

$$\det A=\begin{vmatrix}1&2&3\\3&2&7\\1&0&2\end{vmatrix}=\begin{vmatrix}1&2&3\\1&2&3\\1&0&2\end{vmatrix}+2\begin{vmatrix}1&2&3\\1&0&2\\1&0&2\end{vmatrix}=0+2\cdot 0=0.$$

9°. *The determinant will not change if the elements of one of its rows (columns) are summed up with the corresponding elements of some other row (column) multiplied by the same number.*

For example, $\det A=\begin{vmatrix}2&7&9\\1&3&4\\1&0&2\end{vmatrix}=-1$ (see the example which illustrates property 7°). We multiply the third row by 3 and sum up with the second row. We obtain

$$\det A=\begin{vmatrix}2&7&9\\4&3&10\\1&0&2\end{vmatrix}$$

$$=2\cdot 3\cdot 2+1\cdot 7\cdot 10+4\cdot 0\cdot 9-1\cdot 3\cdot 9-4\cdot 7\cdot 2-2\cdot 0\cdot 10=-1.$$

Using these properties of a determinant, we can simplify the calculations of an nth-order determinant. Transformations which do not alter the value of a determinant are said to be *elementary*.

Example 6. Calculate the determinant

$$d=\begin{vmatrix}1&1&1&1\\-1&2&3&4\\-2&3&6&10\\3&4&10&20\end{vmatrix},$$

expanding it according to the elements of the first column.

$\triangle$ According to formula (5), $d=a_{11}A_{11}+a_{21}A_{21}+a_{31}A_{31}+a_{41}A_{41}$. We seek the cofactors. We have

$$A_{11}=(-1)^{1+1}=\begin{vmatrix}2&3&4\\3&6&10\\4&10&20\end{vmatrix}=2\cdot 2\begin{vmatrix}2&3&2\\3&6&5\\2&5&5\end{vmatrix}$$

$$=4\left(2\begin{vmatrix}6&5\\5&5\end{vmatrix}-3\begin{vmatrix}3&2\\5&5\end{vmatrix}+2\begin{vmatrix}3&2\\6&5\end{vmatrix}\right)=4\,(2\cdot 5-3\cdot 5+2\cdot 3)=4.$$

We have put the common multiple 2 of the third row and the common multiple 2 of the third column before the sign of the determinant A_{11} and then expanded the resulting determinant according to the elements of the first column.

We calculate

$$A_{21}=(-1)^{2+1}\begin{vmatrix}1&1&1\\3&6&10\\4&10&20\end{vmatrix}=-2\begin{vmatrix}1&1&1\\3&6&10\\2&5&10\end{vmatrix}$$

$$=-2\left(\begin{vmatrix}6&10\\5&10\end{vmatrix}-\begin{vmatrix}3&10\\2&10\end{vmatrix}+\begin{vmatrix}3&6\\2&5\end{vmatrix}\right)=-2\,(10-10+3)=-6.$$

We have put the common multiple 2 of the third row before the sign of the determinant and expanded the determinant according to the elements of the first row.

By analogy we obtain

$$A_{31}=(-1)^{3+1}\begin{vmatrix}1&1&1\\2&3&4\\4&10&20\end{vmatrix}=2\begin{vmatrix}1&1&1\\2&3&4\\2&5&10\end{vmatrix}$$

$$=2\left(\begin{vmatrix}3&4\\5&10\end{vmatrix}-\begin{vmatrix}2&4\\2&10\end{vmatrix}+\begin{vmatrix}2&3\\2&5\end{vmatrix}\right)$$

$$=2\,(10-12+4)=4;$$

$$A_{41}=(-1)^{4+1}\begin{vmatrix}1&1&1\\2&3&4\\3&6&10\end{vmatrix}$$

$$=-\left(\begin{vmatrix}3&4\\6&10\end{vmatrix}-\begin{vmatrix}2&4\\3&10\end{vmatrix}+\begin{vmatrix}2&3\\3&6\end{vmatrix}\right)=-6+8-3=-1.$$

Expanding the determinant A according to the elements of the first column, we finally obtain

$$d=1\cdot4+(-1)\cdot(-6)+(-2)\cdot4+3\cdot(-1)=-1.\ \blacktriangle$$

We can considerably simplify the calculation of a determinant if, using the properties of the determinant, we transform it so that formula (6) can be used in calculations.

Example 7. Calculate the determinant

$$d=\begin{vmatrix}1&1&1&1\\-1&2&3&4\\-2&3&6&10\\3&4&10&20\end{vmatrix}$$

using elementary transformations.

$\triangle$ Using the elementary transformations of a determinant, we turn all the elements of the first row, except for $a_{11}=1$, into zero. For that purpose, without changing the first column, we multiply all of its elements by (—1) and add successively to the second, the third and the fourth column. We obtain

$$d=\begin{vmatrix}1&1&1&1\\-1&2&3&4\\-2&3&6&10\\3&4&10&20\end{vmatrix}=\begin{vmatrix}1&0&0&0\\1&3&4&5\\1&5&8&12\\1&1&7&17\end{vmatrix}$$

$$=1\cdot(-1)^{1+1}\begin{vmatrix}3&4&5\\5&8&12\\1&7&17\end{vmatrix}.$$

In the resulting third-order determinant we again turn all of the elements of the third row, except for the first, into zero. For that purpose, without changing the first column, we multiply it consequently by (—7) and (—17) and add successively to the second and the third column. We expand the resulting determinant according to the elements of the third row:

$$d=\begin{vmatrix}3&-17&-46\\5&-27&-73\\1&0&0\end{vmatrix}=1\cdot(-1)^{3+1}\begin{vmatrix}-17&-46\\-27&-73\end{vmatrix}=-1. \blacktriangle$$

2.4. The Inverse Matrix

A square matrix is said to be the *inverse* of a given square matrix if its premultiplication and postmultiplication by the given matrix yield an identity matrix. The inverse of the matrix A is designated as A^{-1}. By definition,

$$AA^{-1}=A^{-1}A=I. \tag{1}$$

A square matrix is said to be *nonsingular* (*invertible*) if its determinant is nonzero. Now if the determinant of the matrix is zero, then the matrix is *singular* (or *noninvertible*).

Theorem. *For the square matrix A to have an inverse, it is necessary and sufficient that the determinant of the matrix A be nonzero, i.e. that the matrix A be nonsingular.*

The process of finding an inverse of a matrix is known as the *inversion* of a matrix.

Let us consider the process of the inversion of a matrix. Let

$$A=\begin{bmatrix}a_{11}&a_{12}&\cdots&a_{1n}\\a_{21}&a_{22}&\cdots&a_{2n}\\\cdot&\cdot&\cdot&\cdot\\a_{n1}&a_{n2}&\cdots&a_{nn}\end{bmatrix} \tag{2}$$

be a nonsingular square matrix of order n whose determinant $d \neq 0$. We compose a matrix of the cofactors of the elements of the given matrix and then transpose it. The resulting matrix is known as an *adjoint*, or *conjugate*, matrix with respect to the matrix A and is designated as $\tilde{A}$:

$$\tilde{A} = \begin{bmatrix} A_{11} & A_{21} & \dots & A_{n1} \\ A_{12} & A_{22} & \dots & A_{n2} \\ \cdot & \cdot & \cdot & \cdot \\ A_{n1} & A_{n2} & \dots & A_{nn} \end{bmatrix}. \tag{3}$$

Calculating the products $A\tilde{A}$ and $\tilde{A}A$ according to the rules of multiplication of matrices, we obtain

$$A\tilde{A} = \tilde{A}A = dI. \tag{4}$$

☐ We shall prove the validity of equality (4) using the example of a third-order matrix. Assume that

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \quad \text{and} \quad \tilde{A} = \begin{bmatrix} A_{11} & A_{21} & A_{31} \\ A_{12} & A_{22} & A_{32} \\ A_{13} & A_{23} & A_{33} \end{bmatrix}.$$

Then

$$A\tilde{A} = \begin{bmatrix} a_{11}A_{11} + a_{12}A_{12} + a_{13}A_{13} & a_{11}A_{21} + a_{12}A_{22} + a_{13}A_{23} & a_{11}A_{31} + a_{12}A_{32} + a_{13}A_{33} \\ a_{21}A_{11} + a_{22}A_{12} + a_{23}A_{13} & a_{21}A_{21} + a_{22}A_{22} + a_{23}A_{23} & a_{21}A_{31} + a_{22}A_{32} + a_{23}A_{33} \\ a_{31}A_{11} + a_{32}A_{12} + a_{33}A_{13} & a_{31}A_{21} + a_{32}A_{22} + a_{33}A_{23} & a_{31}A_{31} + a_{32}A_{32} + a_{33}A_{33} \end{bmatrix}.$$

In accordance with Theorem 3 from 2.3, all the elements of the product $A\tilde{A}$, except for the diagonal ones, are zero. Consequently,

$$A\tilde{A} = \begin{bmatrix} d & 0 & 0 \\ 0 & d & 0 \\ 0 & 0 & d \end{bmatrix} = d \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = dI.$$

We can show by analogy that $A\tilde{A} = dI$. ■

Since $A\tilde{A} = \tilde{A}A = dI$, with $d \neq 0$, it follows that

$$A\frac{\tilde{A}}{d} = \frac{\tilde{A}}{d}A = I.$$

On the other hand, by the definition of the inverse matrix, we have

$$AA^{-1} = A^{-1}A = I.$$

Comparing the last matrix equalities, we get a formula for seeking the inverse matrix:

$$A^{-1} = \tilde{A}/d = \begin{bmatrix} A_{11}/d & A_{21}/d & A_{31}/d \\ A_{12}/d & A_{22}/d & A_{32}/d \\ A_{13}/d & A_{23}/d & A_{33}/d \end{bmatrix}.$$

In the general form, the inverse of a nonsingular square matrix of order n can be calculated by the formula

$$A^{-1} = \begin{bmatrix} A_{11}/d & A_{21}/d & \dots & A_{n1}/d \\ A_{12}/d & A_{22}/d & \dots & A_{n2}/d \\ A_{1n}/d & A_{2n}/d & \dots & A_{nn}/d \end{bmatrix}, \tag{5}$$

i.e. the elements of the original and the inverse matrix are related as $a_{ij}^{-1} = A_{ji}/d$.

Example 1. Find the inverse matrix A^{-1} of the matrix

$$A = \begin{bmatrix} 2 & 3 & 3 \\ 1 & 2 & 3 \\ 2 & 4 & 5 \end{bmatrix}.$$

(1) We calculate the determinant of the matrix A:

$$d = \begin{vmatrix} 2 & 3 & 3 \\ 1 & 2 & 3 \\ 2 & 4 & 5 \end{vmatrix} = 2\cdot2\cdot5 + 3\cdot3\cdot2 + 1\cdot4\cdot3 - 2\cdot2\cdot3 - 1\cdot3\cdot5 - 4\cdot3\cdot2 = -1.$$

Since $d \neq 0$, the inverse matrix A^{-1} exists.

(2) We find the cofactors of the elements of the matrix A:

$$A_{11} = \begin{vmatrix} 2 & 3 \\ 4 & 5 \end{vmatrix} = -2, \quad A_{21} = -\begin{vmatrix} 3 & 3 \\ 4 & 5 \end{vmatrix} = -3, \quad A_{31} = \begin{vmatrix} 3 & 3 \\ 2 & 3 \end{vmatrix} = 3,$$

$$A_{12} = -\begin{vmatrix} 1 & 3 \\ 2 & 5 \end{vmatrix} = 1, \quad A_{22} = \begin{vmatrix} 2 & 3 \\ 2 & 5 \end{vmatrix} = 4, \quad A_{32} = -\begin{vmatrix} 2 & 3 \\ 1 & 3 \end{vmatrix} = -3,$$

$$A_{13} = \begin{vmatrix} 1 & 2 \\ 2 & 4 \end{vmatrix} = 0, \quad A_{23} = -\begin{vmatrix} 2 & 3 \\ 2 & 4 \end{vmatrix} = -2, \quad A_{33} = \begin{vmatrix} 2 & 3 \\ 1 & 2 \end{vmatrix} = 1.$$

(3) We compose an adjoint matrix

$$\tilde{A} = \begin{bmatrix} A_{11} & A_{21} & A_{31} \\ A_{12} & A_{22} & A_{32} \\ A_{13} & A_{23} & A_{33} \end{bmatrix} = \begin{bmatrix} -2 & -3 & 3 \\ 1 & 4 & -3 \\ 0 & -2 & 1 \end{bmatrix}.$$

(4) We calculate the inverse matrix

$$A^{-1}=\widetilde{A}/d\begin{bmatrix} 2 & 3 & -3\\ -1 & -4 & 3\\ 0 & 2 & -1\end{bmatrix}.$$

Verification: $AA^{-1}=\begin{bmatrix} 2 & 3 & 3\\ 1 & 2 & 3\\ 2 & 4 & 5\end{bmatrix}\begin{bmatrix} 2 & 3 & -3\\ -1 & -4 & 3\\ 0 & 2 & -1\end{bmatrix}=\begin{bmatrix} 1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$. ▲

Example 2. Inverse the matrix

$$A=\begin{bmatrix} 1 & 2 & 4 & 2\\ 3 & 1 & 1 & -3\\ -2 & 3 & -1 & 1\\ -1 & 2 & 0 & 1\end{bmatrix}.$$

△ (1) We calculate

$$d=\begin{vmatrix} 1 & 2 & 4 & 2\\ 3 & 1 & 1 & -3\\ -2 & 3 & -1 & 1\\ -1 & 2 & 0 & 1\end{vmatrix},$$

for which purpose we expand the determinant according to the elements of the first row. We have

$$A_{11}=\begin{vmatrix} 1 & 1 & -3\\ 3 & -1 & 1\\ 2 & 0 & 1\end{vmatrix}=-1+2-6-3=-8.$$

$$A_{12}=-\begin{vmatrix} 3 & 1 & -3\\ -2 & -1 & 1\\ -1 & 0 & 1\end{vmatrix}=-(-3-1+3+2)=-1.$$

$$A_{13}=\begin{vmatrix} 3 & 1 & -3\\ -2 & 3 & 1\\ -1 & 2 & 1\end{vmatrix}=9-1+12-9+2-6=7.$$

$$A_{14}=-\begin{vmatrix} 3 & 1 & 1\\ -2 & 3 & -1\\ -1 & 2 & 0\end{vmatrix}=-(1-4+3+6)=-6.$$

Consequently,

$$d=1\cdot(-8)+2\cdot(-1)+4\cdot 7+2\cdot(-6)=6,$$

i.e. the inverse matrix exists.

(2) We calculate the other cofactors of the elements of the matrix A:

$$A_{21}=-\begin{vmatrix} 2 & 4 & 2\\ 3 & -1 & 1\end{vmatrix}=-(-2+8+4-12)=2;$$

$$A_{22}=\begin{vmatrix}1 & 4 & 2\\ -2 & -1 & 1\\ -1 & 0 & 1\end{vmatrix}=-1-4-2+8=1;$$

$$A_{23}=-\begin{vmatrix}1 & 2 & 2\\ -2 & 3 & 1\\ -1 & 2 & 1\end{vmatrix}=-(3-2-8+6+4-2)=-1;$$

$$A_{24}=\begin{vmatrix}1 & 2 & 4\\ -2 & 3 & -1\\ -1 & 2 & 0\end{vmatrix}=2-16+12+2=0;$$

$$A_{31}=\begin{vmatrix}2 & 4 & 2\\ 1 & 1 & -3\\ 2 & 0 & 1\end{vmatrix}=2-24-4-4=-30;$$

$$A_{32}=-\begin{vmatrix}1 & 4 & 2\\ 3 & 1 & -3\\ -1 & 0 & 1\end{vmatrix}=-(1+12+2-12)=-3;$$

$$A_{33}=\begin{vmatrix}1 & 2 & 2\\ 3 & 1 & -3\\ -1 & 2 & 1\end{vmatrix}=1+6+12+2+6-6=21;$$

$$A_{34}=-\begin{vmatrix}1 & 2 & 4\\ 3 & 1 & 1\\ -1 & 2 & 0\end{vmatrix}=-(-2+24+4-2)=-24;$$

$$A_{41}=-\begin{vmatrix}2 & 4 & 2\\ 1 & 1 & -3\\ 3 & -1 & 1\end{vmatrix}=-(2-36-2-6-6-4)=52;$$

$$A_{42}=\begin{vmatrix}1 & 4 & 2\\ 3 & 1 & -3\\ -2 & -1 & 1\end{vmatrix}=1+24-6+4-12-3=8;$$

$$A_{43}=-\begin{vmatrix}1 & 2 & 2\\ 3 & 1 & -3\\ -2 & 3 & 1\end{vmatrix}=-(1+12+18+4+9-6)=-38;$$

$$A_{44}=\begin{vmatrix}1 & 2 & 4\\ 3 & 1 & 1\\ -2 & 3 & -1\end{vmatrix}=-1-4+36+8-3+6=42.$$

(3) We compose an adjoint matrix

$$\tilde{A}=\begin{bmatrix}A_{11} & A_{21} & A_{31} & A_{41}\\ A_{12} & A_{22} & A_{32} & A_{42}\\ A_{13} & A_{23} & A_{33} & A_{43}\\ A_{14} & A_{24} & A_{34} & A_{44}\end{bmatrix}=\begin{bmatrix}-8 & 2 & -30 & 52\\ -1 & 1 & -3 & 8\\ 7 & -1 & 21 & -38\\ -6 & 0 & -24 & 42\end{bmatrix}.$$

(4) We divide all of the elements of the adjoint matrix by $d = 6$ and obtain an inverse matrix A^{-1}:

$$A^{-1}=\begin{bmatrix} -8/6 & 2/6 & -30/6 & 52/6 \\ -1/6 & 1/6 & -3/6 & 8/6 \\ 7/6 & -1/6 & 21/6 & -38/6 \\ -6/6 & 0 & -24/6 & 42/6 \end{bmatrix}.$$

Verification:

$$AA^{-1}=\begin{bmatrix} 1 & 2 & 4 & 2 \\ 3 & 1 & 1 & -3 \\ -2 & 3 & -1 & 1 \\ -1 & 2 & 0 & 1 \end{bmatrix}\begin{bmatrix} -8/6 & 2/6 & -30/6 & 52/6 \\ -1/6 & 1/6 & -3/6 & 8/6 \\ 7/6 & -1/6 & 21/6 & -38/6 \\ -6/6 & 0 & -24/6 & 42/6 \end{bmatrix}$$

$$=\begin{bmatrix} \frac{-8-2+28-12}{6} & \frac{2+2-4+0}{6} & \frac{-30-6+84-48}{6} & \frac{52+16-152+84}{6} \\ \frac{-24-1+7+18}{6} & \frac{6+1-1+0}{6} & \frac{-90-3+21+72}{6} & \frac{156+8-38-126}{6} \\ \frac{16-3-7-6}{6} & \frac{-4+3+1+0}{6} & \frac{60-9-21-24}{6} & \frac{-104+24+38+42}{6} \\ \frac{8-2+0-6}{6} & \frac{-2+2+0+0}{6} & \frac{30-6+0-24}{6} & \frac{-52+16+0+42}{6} \end{bmatrix}$$

$$=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}=I. \blacktriangle$$

If the order of the matrix A is high, then this method of inversion of a matrix is very laborious. There are other methods of inversion of a matrix which we shall consider later on.

It is very significant to find the inverse matrix A^{-1} for solving systems of linear equations.

2.5. Solving Matrix Equations

We shall consider three kinds of matrix equations

$$AX = B, \tag{1}$$

$$XA = B, \tag{2}$$

$$AXB = C, \tag{3}$$

where

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \cdot & \cdot\ \cdot\ \cdot & \cdot \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}, \quad B = \begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \cdot & \cdot\ \cdot\ \cdot & \cdot \\ b_{n1} & \cdots & b_{nn} \end{bmatrix},$$

$$C = \begin{bmatrix} c_{11} & \cdots & c_{1n} \\ \cdot & \cdot\ \cdot\ \cdot & \cdot \\ c_{n1} & \cdots & c_{nn} \end{bmatrix}$$

are specified square matrices of the same dimension, and

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \cdot & \cdot\ \cdot\ \cdot & \cdot \\ x_{n1} & \cdots & x_{nn} \end{bmatrix}$$

is a square matrix of the same dimension whose elements are unknown.

Let us solve each of the equations (1)-(3).

To solve equation (1), we premultiply two its sides by A^{-1} (on the assumption that the inverse matrix A^{-1} exists):

$$A^{-1}AX = A^{-1}B.$$

But the product $A^{-1}A = I$ and, consequently, $IX = A^{-1}B$, whence

$$X = A^{-1}B. \tag{4}$$

Example 1. Solve the matrix equation

$$\underbrace{\begin{bmatrix} 2 & 5 \\ -1 & 9 \end{bmatrix}}_{A} X = \underbrace{\begin{bmatrix} -2 & 3 \\ 1 & -1 \end{bmatrix}}_{B}.$$

△ (1) $\det A = \begin{vmatrix} 2 & 5 \\ -1 & 9 \end{vmatrix} = 23 \neq 0.$

We seek A^{-1}. Since $A_{11} = 9$, $A_{21} = -5$, $A_{12} = 1$, $A_{22} = 2$, follows that

$$A = \begin{bmatrix} 9 & -5 \\ 1 & 2 \end{bmatrix}, \text{ whence } A^{-1} = \frac{1}{23}\begin{bmatrix} 9 & -5 \\ 1 & 2 \end{bmatrix}.$$

(3) From formula (4) we find that

$$X = A^{-1}B = \frac{1}{23}\begin{bmatrix} 9 & -5 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} -2 & 3 \\ 1 & -1 \end{bmatrix} = \frac{1}{23}\begin{bmatrix} -23 & 22 \\ 0 & 5 \end{bmatrix}. \ \blacktriangle$$

In practical calculations we often encounter equations of form (1), where **x** and **b** are vector columns of the same dimension as the matrix A.

Example 2. Solve the matrix equation

$$\underbrace{\begin{bmatrix} 1 & 1 & 2 \\ 2 & -1 & 2 \\ 4 & 1 & 4 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}}_{\mathbf{x}} = \underbrace{\begin{bmatrix} -1 \\ -4 \\ -2 \end{bmatrix}}_{\mathbf{b}}.$$

△ (1) $\det A = \begin{vmatrix} 1 & 1 & 2 \\ 2 & -1 & 2 \\ 4 & 1 & 4 \end{vmatrix} = -4 + 8 + 4 + 8 - 2 - 8 = 6 \neq 0.$

(2) We seek A^{-1}. We have $A_{11} = -6$, $A_{21} = -2$, $A_{31} = 4$, $A_{12} = 0$, $A_{22} = -4$, $A_{32} = 2$, $A_{13} = 6$, $A_{23} = 3$, $A_{33} = -3$, i.e.

$$\tilde{A} = \begin{bmatrix} -6 & -2 & 4 \\ 0 & -4 & 2 \\ 6 & 3 & -3 \end{bmatrix}, \quad A^{-1} = \begin{bmatrix} -1 & -1/3 & 2/3 \\ 0 & -2/3 & 1/3 \\ 1 & 1/2 & -1/2 \end{bmatrix}.$$

(3) From formula (4) we obtain

$$\mathbf{x} = A^{-1}\mathbf{b} = \begin{bmatrix} -1 & -1/3 & 2/3 \\ 0 & -2/3 & 1/3 \\ 1 & 1/2 & -1/2 \end{bmatrix}\begin{bmatrix} -1 \\ -4 \\ -2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ -2 \end{bmatrix}. \ \blacktriangle$$

To solve equation (2), we postmultiply its two sides by A^{-1} (on the assumption that the inverse matrix A^{-1} exists):

$$XAA^{-1} = BA^{-1}.$$

This means that $XI = BA^{-1}$, whence it follows that

$$X = BA^{-1}. \qquad (5)$$

Example 3. Solve the matrix equation

$$X\underbrace{\begin{bmatrix} 1 & 1 & -1 \\ 2 & -1 & 1 \\ 1 & 0 & 1 \end{bmatrix}}_{A}=\underbrace{\begin{bmatrix} 6 & 2 & -1 \\ 6 & 1 & 1 \\ 8 & -1 & 4 \end{bmatrix}}_{B}.$$

△ We find that

$$\det A=\begin{vmatrix} 1 & 1 & -1 \\ 2 & -1 & 1 \\ 1 & 0 & 1 \end{vmatrix}=-3,\ A_{11}=-1,\quad A_{21}=-1,\ A_{31}=0,$$

$$A_{12}=-1,\ A_{22}=2,\ A_{32}=-3,\ A_{13}=1,\ A_{23}=1,\ A_{33}=-3,$$

$$\tilde{A}=\begin{bmatrix} -1 & -1 & 0 \\ -1 & 2 & -3 \\ 1 & 1 & -3 \end{bmatrix},\quad A^{-1}=\begin{bmatrix} 1/3 & 1/3 & 0 \\ 1/3 & -2/3 & 1 \\ -1/3 & -1/3 & 1 \end{bmatrix}.$$

From formula (5) we obtain

$$X=BA^{-1}\begin{bmatrix} 6 & 2 & -1 \\ 6 & 1 & 1 \\ 8 & -1 & 4 \end{bmatrix}\begin{bmatrix} 1/3 & 1/3 & 0 \\ 1/3 & -2/3 & 1 \\ -1/3 & -1/3 & 1 \end{bmatrix}=\begin{bmatrix} 3 & 1 & 1 \\ 2 & 1 & 2 \\ 1 & 2 & 3 \end{bmatrix}.$$

To solve equation (3), we premultiply both its sides by A^{-1} and postmultiply them by B^{-1} (on the assumption that the indicated inverse matrices exist). Then we obtain

$$A^{-1}AXBB^{-1}=A^{-1}CB^{-1},\ \text{or}\ IXI=A^{-1}CB^{-1}$$

whence if follows that

$$X=A^{-1}CB^{-1}. \tag{6}$$

Example 4. Solve the matrix equation

$$\underbrace{\begin{bmatrix} 1 & -3 & 2 \\ 3 & -4 & 0 \\ 2 & -5 & 3 \end{bmatrix}}_{A}X\underbrace{\begin{bmatrix} 3 & 1 & 1 \\ 2 & 1 & 2 \\ 1 & 2 & 3 \end{bmatrix}}_{B}=\underbrace{\begin{bmatrix} 8 & -4 & -4 \\ 18 & 5 & 10 \\ 17 & -3 & -1 \end{bmatrix}}_{C}.$$

△ We seek

$$A^{-1}=\begin{bmatrix} -12 & -1 & 8 \\ -9 & -1 & 6 \\ -7 & -1 & 5 \end{bmatrix},\quad B^{-1}=\begin{bmatrix} 1 & 1/4 & -1/4 \\ 1 & -2 & 1 \\ -3/4 & 5/4 & -1/4 \end{bmatrix}$$

(we recommend the reader to carry out the calculations independently). Next we have

$$A^{-1}C=\begin{bmatrix}-12 & -1 & 8\\ -9 & -1 & 6\\ -7 & -1 & 5\end{bmatrix}\begin{bmatrix}8 & -4 & -4\\ 18 & 5 & 10\\ 17 & -3 & -1\end{bmatrix}=\begin{bmatrix}22 & 19 & 30\\ 12 & 13 & 20\\ 11 & 8 & 13\end{bmatrix}.$$

Now we find from formula (6) that

$$X=A^{-1}CB^{-1}=\begin{bmatrix}22 & 19 & 30\\ 12 & 13 & 20\\ 11 & 8 & 13\end{bmatrix}\begin{bmatrix}1/4 & 1/4 & -1/4\\ 1 & -2 & 1\\ -3/4 & 5/4 & -1/4\end{bmatrix}=\begin{bmatrix}2 & 5 & 6\\ 1 & 2 & 5\\ 1 & 3 & 2\end{bmatrix}. \blacktriangle$$

2.6. Triangular Matrices. Expansion of a Matrix in a Product of Two Triangular Matrices

A square matrix is said to be *triangular* if the elements which are higher or lower than the principal diagonal are zero. If the elements which are higher than the principal diagonal are zero, then the matrix is a *lower triangular*, or *subdiagonal*, matrix. Such is, for example, the matrix

$$T_1=\begin{bmatrix}t_{11} & 0 & \dots & 0\\ t_{21} & t_{22} & \dots & 0\\ \cdot & \cdot & \cdot & \cdot\\ t_{n1} & t_{n2} & \dots & t_{nn}\end{bmatrix}.$$

Now if the elements which are lower than the principal diagonal are zero, then the matrix is an *upper triangular*, or *superdiagonal*, matrix. Such is, for example, the matrix

$$T_2=\begin{bmatrix}r_{11} & r_{12} & \dots & r_{1n}\\ 0 & r_{22} & \dots & r_{2n}\\ \cdot & \cdot & \cdot & \cdot\\ 0 & 0 & \dots & r_{nn}\end{bmatrix}.$$

The determinant of a triangular matrix is equal to the product of its diagonal elements. If $T=[t_{ij}]$ is a triangular matrix, then

$$\det T=t_{11}t_{22}\dots t_{nn}.$$

The inverse of a nonsingular triangular matrix is also a triangular matrix of the same dimension and structure.

If the square matrix

$$A=\begin{bmatrix}a_{11} & a_{12} & \dots & a_{1n}\\ a_{21} & a_{22} & \dots & a_{2n}\\ \cdot & \cdot & \cdot & \cdot\\ a_{n1} & a_{n2} & \dots & a_{nn}\end{bmatrix}$$

has nonzero *diagonal minors* (this is the name for the minors of the determinant of the matrix which have the diagonal elements of the matrix on their principal diagonals), i.e.

$$a_{11} \neq 0, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \neq 0, \quad \ldots, \quad \det A \neq 0,$$

then it can be expanded in the product of two triangular matrices (the upper and the lower). This expansion is unique if nonzero values are preassigned to the diagonal elements of one of the triangular matrices (say, they are set equal to unity).

Assume that

$$A = CB,$$

where C is a lower triangular matrix and B is an upper triangular matrix with diagonal elements equal to unity.

Using a fourth-order matrix as an example, we shall obtain formulas which express the relationship between the elements of the matrices A, B and C:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} c_{11} & 0 & 0 & 0 \\ c_{21} & c_{22} & 0 & 0 \\ c_{31} & c_{32} & c_{33} & 0 \\ c_{41} & c_{42} & c_{43} & c_{44} \end{bmatrix} \begin{bmatrix} 1 & b_{12} & b_{13} & b_{14} \\ 0 & 1 & b_{23} & b_{24} \\ 0 & 0 & 1 & b_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

We multiply and then equate the resulting elements of the matrix CB to the respective elements of the matrix A:

$$c_{11} = a_{11}, \tag{1}$$
$$c_{21} = a_{21}, \tag{2}$$
$$c_{31} = a_{31}, \tag{3}$$
$$c_{41} = a_{41}, \tag{4}$$
$$c_{11}b_{12} = a_{12}, \tag{5}$$
$$c_{21}b_{12} + c_{22} = a_{22}, \tag{6}$$
$$c_{31}b_{12} + c_{32} = a_{32}, \tag{7}$$
$$c_{41}b_{12} + c_{42} = a_{42}, \tag{8}$$
$$c_{11}b_{13} = a_{13}, \tag{9}$$
$$c_{21}b_{13} + c_{22}b_{23} = a_{23}, \tag{10}$$
$$c_{31}b_{13} + c_{32}b_{23} + c_{33} = a_{33}, \tag{11}$$

$$c_{41}b_{13} + c_{42}b_{23} + c_{43} = a_{43}, \qquad (12)$$

$$c_{11}b_{14} = a_{14}, \qquad (13)$$

$$c_{21}b_{14} + c_{22}b_{24} = a_{24}, \qquad (14)$$

$$c_{31}b_{14} + c_{32}b_{24} + c_{33}b_{34} = a_{34}, \qquad (15)$$

$$c_{41}b_{14} + c_{42}b_{24} + c_{43}b_{34} + c_{44} = a_{44}. \qquad (16)$$

From equations (1)-(16) we find the elements b_{ij} and c_{ij} $(i = 1, 2, 3, 4, j = 1, 2, 3, 4)$ of the triangular matrices B and C in the following order:

(I) the first column of the matrix C [formulas (1)-(4)]:

$$c_{i1} = a_{i1}, \quad i = 1, 2, 3, 4.$$

(II) the first row of the matrix B [formulas (5), (9), (13)]:

$$b_{1j} = a_{1j}/c_{11}, \quad j = 2, 3, 4.$$

(III) the second column of the matrix C [formulas (6), (7), (8)]:

$$c_{i2} = a_{i2} - c_{i1}b_{12}, \quad i = 2, 3, 4.$$

(IV) the second row of the matrix B [formulas (10), (14)]:

$$b_{2j} = (a_{2j} - c_{21}b_{1j})/c_{22}, \quad j = 3, 4.$$

(V) the third column of the matrix C [formulas (11), (12)]:

$$c_{i3} = a_{i3} - c_{i1}b_{13} - c_{i2}b_{23}, \; i = 3, 4.$$

(VI) the third row of the matrix B [formula (15)]:

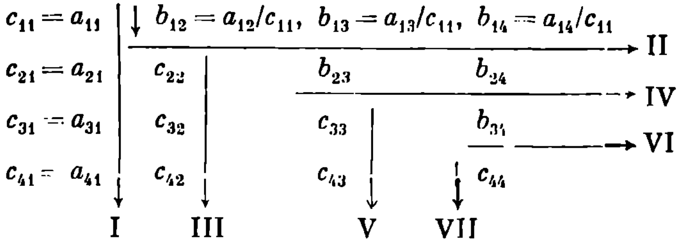
$$b_{34} = (a_{34} - c_{31}b_{14} - c_{32}b_{24})/c_{33}.$$

(VII) the fourth column of the matrix C [formula (16)]:

$$c_{44} = a_{44} - c_{41}b_{14} - c_{42}b_{24} - c_{43}b_{34}.$$

Scheme of Successive Determination of the Elements b_{ij} and c_{ij}

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} c_{11} & 0 & 0 & 0 \\ c_{21} & c_{22} & 0 & 0 \\ c_{31} & c_{32} & c_{33} & 0 \\ c_{41} & c_{42} & c_{43} & c_{44} \end{bmatrix} \begin{bmatrix} 1 & b_{12} & b_{13} & b_{14} \\ 0 & 1 & b_{23} & b_{24} \\ 0 & 0 & 1 & b_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$



The roman figures at the arrows show the sequence in which the elements c_{ij} and b_{ij} must be found. In this expansion we first find the columns and then the rows.

To simplify the calculations, it is more convenient to expand the matrix A in the product of two triangular matrices C and B using Table 2.1 given below.

Table 2.1

a_{11}		a_{12}	a_{13}	a_{14}
a_{21}		a_{22}	a_{23}	a_{24}
a_{31}		a_{32}	a_{33}	a_{34}
a_{41}		a_{42}	a_{43}	a_{44}
$c_{11}=a_{11}$	1	$b_{12}=a_{12}/c_{11}$	$b_{13}=a_{13}/c_{11}$	$b_{14}=a_{14}/c_{11}$
$c_{21}=a_{21}$		c_{22} 1	b_{23}	b_{24}
$c_{31}=a_{31}$		c_{32}	c_{33} 1	b_{34}
$c_{41}=a_{41}$		c_{42}	c_{43}	c_{44} 1

This table is composed as follows:

1. On the basis of the formulas indicated above, we write the elements of the first column of the matrix A

in the first column of the matrix C and the elements of the first row of the matrix A, divided by c_{11}, in the first row of the matrix B.

2. The elements which are under the stepped line can be found as follows: we take the requisite element of the matrix A and subtract from it the product of the elements which are to the left in the same row and higher in the same column as the required element, and here we multiply the first element of the row by the first element of the column, the second element of the row by the second element of the column and so on.

For example, $c_{33} = a_{33} - c_{31}b_{13} - c_{32}b_{23}$.

3. When we calculate the elements which are above the stepped line, we do the same as we did in item 2 but divide the result by the diagonal element c_{ii} $(i = 2, 3)$ which is in the same row as the required element.

For example $b_{34} = \frac{a_{34} - c_{31}b_{14} - c_{32}b_{24}}{c_{33}}$.

By analogy we can expand a square matrix of any order n in the product of two triangular matrices. Somewhat higher we have indicated the rule of transformation of a matrix into the product of two triangular matrices for the case $b_{ii} = 1$. Now if $c_{ii} = 1$, then we must first calculate the elements of the rows of the matrix B using the formula

$$b_{ij} = a_{ij} - \sum_{k=1}^{i-1} c_{ik}b_{kj} \; (i \leqslant j), \tag{17}$$

and then the elements of the columns of the matrix C using the formula

$$c_{ij} = \frac{a_{ij} - \sum\limits_{k=1}^{i-1} c_{ik}b_{kj}}{b_{ii}} \; (i > j). \tag{18}$$

We represented the matrix A as the product CB of two triangular matrices, where C is the lower and B is the upper triangular matrix. This order of the factors is not obligatory, however, i.e. we can represent the matrix A as the product BC and obtain similar formulas for the elements of the triangular matrices B and C.

Example. Expand the matrix

$$A=\begin{bmatrix} 1 & 2 & 3 & 4 \\ -1 & 2 & 4 & -3 \\ 2 & 4 & 5 & -2 \\ 4 & 3 & 2 & 1 \end{bmatrix}$$

in the product CB, where

$$C=\begin{bmatrix} c_{11} & 0 & 0 & 0 \\ c_{21} & c_{22} & 0 & 0 \\ c_{31} & c_{32} & c_{33} & 0 \\ c_{41} & c_{42} & c_{43} & c_{44} \end{bmatrix},\quad B=\begin{bmatrix} 1 & b_{12} & b_{13} & b_{14} \\ 0 & 1 & b_{23} & b_{24} \\ 0 & 0 & 1 & b_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

△ The solution is given in Table 2.2.

Table 2.2

<table>
<tr><td>1</td><td colspan="2">2</td><td colspan="2">3</td><td colspan="2">4</td><td rowspan="4">a_{ij}</td></tr>
<tr><td>−1</td><td colspan="2">2</td><td colspan="2">4</td><td colspan="2">−3</td></tr>
<tr><td>2</td><td colspan="2">4</td><td colspan="2">5</td><td colspan="2">−2</td></tr>
<tr><td>4</td><td colspan="2">3</td><td colspan="2">2</td><td colspan="2">1</td></tr>
<tr><td>1</td><td>1</td><td>$\frac{2}{1}=2$</td><td colspan="2">$\frac{3}{1}=3$</td><td colspan="2">$\frac{4}{1}=4$</td><td rowspan="4">b_{ij}</td></tr>
<tr><td>−1</td><td>$2-(-1)\cdot 2=4$</td><td>1</td><td colspan="2">$\frac{4-(-1)\cdot 3}{4}=\frac{7}{4}$</td><td colspan="2">$\frac{-3-(-1)\cdot 4}{4}=\frac{1}{4}$</td></tr>
<tr><td>2</td><td colspan="2">$4-2\cdot 2=0$</td><td>$5-2\cdot 3-0\cdot\frac{7}{4}=-1$</td><td>1</td><td colspan="2">$\frac{-2-2\cdot 4-0\cdot\frac{1}{4}}{-1}=10$</td></tr>
<tr><td>4</td><td colspan="2">$3-4\cdot 2=-5$</td><td colspan="2">$2-4\cdot 3-(-5)\cdot\frac{7}{4}=-\frac{5}{4}$</td><td>$1-4\cdot 4-(-5)\times\frac{1}{4}-\left(-\frac{5}{4}\right)\times 10=-\frac{5}{4}$</td><td>1</td></tr>
</table>

Verification:

$$CB=\begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 4 & 0 & 0 \\ 2 & 0 & -1 & 0 \\ 4 & -5 & -5/4 & -5/4 \end{bmatrix}\begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 1 & 7/4 & 1/4 \\ 0 & 0 & 1 & 10 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$=\begin{bmatrix} 1 & 2 & 3 & 4 \\ -1 & 2 & 4 & -3 \\ 2 & 4 & 5 & -2 \\ 4 & 3 & 2 & 1 \end{bmatrix}=A. \ \blacktriangle$$

2.7. Inversion of a Matrix by Expanding It in a Product of Two Triangular Matrices

It follows from the definition of the inverse matrix that if

$$A = CB, \tag{1}$$

where all the matrices are nonsingular, we can find the inverse matrix using the formula

$$A^{-1} = B^{-1}C^{-1}. \tag{2}$$

☐ To prove the validity of formula (2), we perform the following transformations.

We premultiply both sides of relation (1) by the matrix C^{-1}:

$$C^{-1}A = C^{-1}CB, \text{ or } C^{-1}A = B. \tag{3}$$

We premultiply both sides of relation (3) by the matrix B^{-1}:

$$B^{-1}C^{-1}A = B^{-1}B, \text{ or } (B^{-1}C^{-1})A = I. \tag{4}$$

We postmultiply both sides of relation (4) by the matrix A^{-1}:

$$(B^{-1}C^{-1})AA^{-1} = A^{-1}, \text{ or } A^{-1} = B^{-1}C^{-1}. \ \blacksquare$$

In formula (2) the inverse matrix A^{-1} is expressed as the product of the inverse matrices B^{-1} and C^{-1}. However, if the matrices B and C are triangular, then, to calculate the matrix A^{-1}, it is not necessary to invert the matrices B and C.

Let us derive formulas for calculating the elements of the inverse matrix A^{-1} using the example of a fourth-order matrix. We designate the elements of the matrix A^{-1} as α_{ij} and the elements of the inverse matrices B^{-1} and C^{-1} as β_{ij} and γ_{ij} respectively. The matrices B and C are of the same kind as those given in 2.6. We postmultiply relation (2) by the matrix C and premultiply it by the matrix B:

$$A^{-1}C = B^{-1}C^{-1}C, \text{ or } A^{-1}C = B^{-1}, \tag{5}$$

$$BA^{-1} = BB^{-1}C^{-1}, \text{ or } BA^{-1} = C^{-1}. \tag{6}$$

Matrices C^{-1} and B^{-1} are triangular matrices of the same kind as the matrices C and B. We write relations (5) and (6) as follows:

$$\begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{bmatrix} \begin{bmatrix} c_{11} & 0 & 0 & 0 \\ c_{21} & c_{22} & 0 & 0 \\ c_{31} & c_{32} & c_{33} & 0 \\ c_{41} & c_{42} & c_{43} & c_{44} \end{bmatrix} = \begin{bmatrix} 1 & \beta_{12} & \beta_{13} & \beta_{14} \\ 0 & 1 & \beta_{23} & \beta_{24} \\ 0 & 0 & 1 & \beta_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{7}$$

$$\begin{bmatrix} 1 & b_{12} & b_{13} & b_{14} \\ 0 & 1 & b_{23} & b_{24} \\ 0 & 0 & 1 & b_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{bmatrix} = \begin{bmatrix} \gamma_{11} & 0 & 0 & 0 \\ \gamma_{21} & \gamma_{22} & 0 & 0 \\ \gamma_{31} & \gamma_{32} & \gamma_{33} & 0 \\ \gamma_{41} & \gamma_{42} & \gamma_{43} & \gamma_{44} \end{bmatrix}. \tag{8}$$

We shall multiply only the rows and columns the elements of whose products in the matrices B^{-1} and C^{-1} are equal to zeros and unities respectively.

We first consider the product $A^{-1}C = B^{-1}$:

$$\beta_{11} = \alpha_{11}c_{11} + \alpha_{12}c_{21} + \alpha_{13}c_{31} + \alpha_{14}c_{41} = 1, \tag{9}$$

$$\beta_{21} = \alpha_{21}c_{11} + \alpha_{22}c_{21} + \alpha_{23}c_{31} + \alpha_{24}c_{41} = 0, \tag{10}$$

$$\beta_{31} = \alpha_{31}c_{11} + \alpha_{32}c_{21} + \alpha_{33}c_{31} + \alpha_{34}c_{41} = 0, \tag{11}$$

$$\beta_{41} = \alpha_{41}c_{11} + \alpha_{42}c_{21} + \alpha_{43}c_{31} + \alpha_{44}c_{41} = 0, \tag{12}$$

$$\beta_{22} = \alpha_{21}\cdot 0 + \alpha_{22}c_{22} + \alpha_{23}c_{32} + \alpha_{24}c_{42} = 1, \tag{13}$$

$$\beta_{32} = \alpha_{31}\cdot 0 + \alpha_{32}c_{22} + \alpha_{33}c_{32} + \alpha_{34}c_{42} = 0, \tag{14}$$

$$\beta_{42} = \alpha_{41}\cdot 0 + \alpha_{42}c_{22} + \alpha_{43}c_{32} + \alpha_{44}c_{42} = 0, \tag{15}$$

$$\beta_{33} = \alpha_{31}\cdot 0 + \alpha_{32}\cdot 0 + \alpha_{33}c_{33} + \alpha_{34}c_{43} = 1, \tag{16}$$

$$\beta_{43} = \alpha_{41}\cdot 0 + \alpha_{42}\cdot 0 + \alpha_{43}c_{33} + \alpha_{44}c_{43} = 0, \tag{17}$$

$$\beta_{44} = \alpha_{41}\cdot 0 + \alpha_{42}\cdot 0 + \alpha_{43}\cdot 0 + \alpha_{44}c_{44} = 1. \tag{18}$$

And now we consider the product $BA^{-1} = C^{-1}$:

$$\gamma_{12} = 1\cdot\alpha_{12} + b_{12}\alpha_{22} + b_{13}\alpha_{32} + b_{14}\alpha_{42} = 0, \tag{19}$$

$$\gamma_{13} = 1\cdot\alpha_{13} + b_{12}\alpha_{23} + b_{13}\alpha_{33} + b_{14}\alpha_{43} = 0, \tag{20}$$

$$\gamma_{14} = 1\cdot\alpha_{14} + b_{12}\alpha_{24} + b_{13}\alpha_{34} + b_{14}\alpha_{44} = 0, \tag{21}$$

$$\gamma_{23} = 0\cdot\alpha_{13} + 1\cdot\alpha_{23} + b_{23}\alpha_{33} + b_{24}\alpha_{43} = 0, \tag{22}$$

$$\gamma_{24} = 0\cdot\alpha_{14} + 1\cdot\alpha_{24} + b_{23}\alpha_{34} + b_{24}\alpha_{44} = 0, \tag{23}$$

$$\gamma_{34} = 0\cdot\alpha_{14} + 0\cdot\alpha_{24} + 1\cdot\alpha_{34} + b_{34}\alpha_{44} = 0. \tag{24}$$

We have to find 16 unknown elements α_{ij} from relations (9)-(24). We have

I. $\alpha_{44} = 1/c_{44}$ [from (18)]
$\alpha_{43} = (1/c_{33})(-\alpha_{44}c_{43})$ [from (17)]
$\alpha_{42} = (1/c_{22})(-\alpha_{43}c_{32} - \alpha_{44}c_{42})$ [from (15)]
$\alpha_{41} = (1/c_{11})(-\alpha_{42}c_{21} - \alpha_{43}c_{31} - \alpha_{44}c_{41})$ [from (12)]

II. $\alpha_{34} = -b_{34}\alpha_{44}$ [from (24)]
$\alpha_{24} = -b_{23}\alpha_{34} - b_{24}\alpha_{44}$ [from (23)]
$\alpha_{14} = -b_{12}\alpha_{24} - b_{13}\alpha_{34} - b_{14}\alpha_{44}$ [from (21)]

III. $\alpha_{33} = (1/c_{33})(1 - \alpha_{34}c_{43})$ [from (16)]
$\alpha_{32} = (1/c_{22})(-\alpha_{33}c_{32} - \alpha_{34}c_{42})$ [from (14)]
$\alpha_{31} = (1/c_{11})(-\alpha_{32}c_{21} - \alpha_{33}c_{31} - \alpha_{34}c_{41})$ [from (11)]

IV. $\alpha_{23} = -b_{23}\alpha_{33} - b_{24}\alpha_{43}$ [from (22)]
$\alpha_{13} = -b_{12}\alpha_{23} - b_{13}\alpha_{33} - b_{14}\alpha_{43}$ [from (20)]

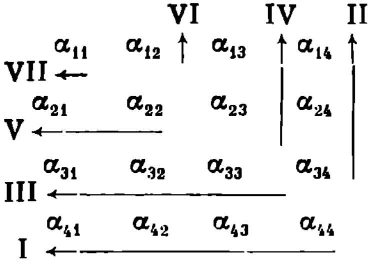
V. $\alpha_{22} = (1/c_{22})(1 - \alpha_{23}c_{32} - \alpha_{24}c_{42})$ [from (13)]
$\alpha_{21} = (1/c_{11})(-\alpha_{22}c_{21} - \alpha_{23}c_{31} - \alpha_{24}c_{41})$ [from (10)]

VI. $\alpha_{12} = -b_{12}\alpha_{22} - b_{13}\alpha_{32} - b_{14}\alpha_{42}$ [from (19)]

VII. $\alpha_{11} = (1/c_{11})(1 - \alpha_{12}c_{21} - \alpha_{13}c_{31} - \alpha_{14}c_{41})$ [from (9)]

Using the relations obtained at stages I-VII, we arrive at the following scheme.

Scheme of Successive Calculations of the Elements of an Inverse Matrix



Using this scheme, we can calculate the inverse matrix A^{-1} of the matrix A of any order if it is expanded in the product CB of two triangular matrices, where C and B are triangular matrices of the same kind as those presented above.

This expansion results in the following formulas for the elements of the matrix A^{-1}:

$$\left.\begin{aligned} &\alpha_{ij} = -\sum_{k=i+1}^{n} b_{ik}\alpha_{kj} \; (i<j), \\ &\alpha_{ii} = \frac{1}{c_{ii}}\left(1-\sum_{k=i+1}^{n}\right)\alpha_{ik}c_{ki} \; (i=j), \\ &\alpha_{ij} = -\frac{\sum\limits_{k=j+1}^{n} \alpha_{ik}c_{kj}}{c_{jj}} \; (i>j). \end{aligned}\right\} \quad (25)$$

Example. Using the expansion of the matrix

$$A=\begin{bmatrix} 3 & -2 & 2 & 0 \\ 2 & 1 & 1 & -2 \\ 3 & -1 & 2 & 1 \\ 1 & 2 & -1 & -1 \end{bmatrix}$$

in the product of two triangular matrices, find the inverse matrix A^{-1}.

△ We compile Table 2.3 (see p. 76).

We calculate the elements of the inverse matrix in the consecutive order indicated above (see the scheme), using formulas (25).

Table 2.3

	3	−2	2	0	A
	2	1	1	−2	
	3	−1	2	1	
	1	2	−1	−1	
C	3	1 −2/3	2/3	0	B
	2	$1-2\left(-\frac{2}{3}\right)=\frac{7}{3}$	1 $\frac{1-2\cdot(2/3)}{7/3}=-\frac{1}{7}$	$\frac{-2-2\cdot 0}{7/3}=-\frac{6}{7}$	
	3	$-1-3\cdot\left(-\frac{2}{3}\right)=1$	$2-3\cdot\frac{2}{3}-1\times\left(-\frac{1}{7}\right)=\frac{1}{7}$	1 $\frac{1-3\cdot 0-1\cdot(-6/7)}{1/7}=13$	
	1	$2-1\cdot\left(-\frac{2}{3}\right)=\frac{8}{3}$	$-1-1\cdot\frac{2}{3}-\frac{8}{3}\times\left(-\frac{1}{7}\right)=-\frac{9}{7}$	$-1-1\cdot 0-\frac{8}{3}\times\left(-\frac{6}{7}\right)-\left(-\frac{8}{7}\right)\times 13=18$ 1	
	1/3	−4/18	0	8/18	A^{-1}
	−12/18	5/18	1/2	−1/18	
	−12/18	11/18	1/2	−13/18	
	−1/3	−5/18	1/2	1/18	

I. $\alpha_{44} = 1/c_{44} = 1/18,$

$$\alpha_{43} = -\frac{\alpha_{44}c_{43}}{c_{33}} = -\frac{(1/18)\cdot(-9/7)}{1/7} = \frac{1}{2},$$

$$\alpha_{42} = -\frac{\alpha_{43}c_{32} + \alpha_{44}c_{42}}{c_{22}} = -\frac{(1/2)\cdot 1 + (1/18)\cdot(8/3)}{7/3} = -\frac{5}{18},$$

$$\alpha_{41} = -\frac{\alpha_{42}c_{21} + \alpha_{43}c_{31} + \alpha_{44}c_{41}}{c_{11}}$$

$$= -\frac{(-5/18)\cdot 2 + (1/2)\cdot 3 + (1/18)\cdot 1}{3} = -\frac{1}{3}.$$

II. $\alpha_{34} = -b_{34}\alpha_{44} = -13/18,$

$$\alpha_{24} = -(b_{23}\alpha_{34} + b_{24}\alpha_{44})$$

$$= -\left[\left(-\frac{1}{7}\right)\cdot\left(-\frac{13}{18}\right) + \left(-\frac{6}{7}\right)\cdot\frac{1}{18}\right] = -\frac{1}{18},$$

$$\alpha_{14} = -(b_{12}\alpha_{24} + b_{13}\alpha_{34} + b_{14}\alpha_{44})$$

$$= -\left[\left(-\frac{2}{3}\right)\left(-\frac{1}{18}\right) + \frac{2}{3}\cdot\left(-\frac{13}{18}\right) + 0\cdot\frac{1}{18}\right] = \frac{8}{18}.$$

III. $$\alpha_{33} = \frac{1}{c_{33}}(1 - \alpha_{34}c_{43}) = \frac{1}{1/7}\left[1 - \left(-\frac{13}{18}\right)\cdot\left(-\frac{9}{7}\right)\right] = \frac{1}{2},$$

$$\alpha_{32} = -\frac{\alpha_{33}c_{32} + \alpha_{34}c_{42}}{c_{22}} = -\frac{(1/2)\cdot 1 + (-13/18)\cdot(8/3)}{7/3} = \frac{11}{18},$$

$$\alpha_{31} = -\frac{\alpha_{32}c_{21} + \alpha_{33}c_{31} + \alpha_{34}c_{41}}{c_{11}}$$

$$= -\frac{(11/18)\cdot 2 + (1/2)\cdot 3 + (-13/18)\cdot 1}{3} = -\frac{12}{18}.$$

IV. $\alpha_{23} = -(b_{23}\alpha_{33} + b_{24}\alpha_{43})$

$$= -\left[\left(-\frac{1}{7}\right)\cdot\frac{1}{2} + \left(-\frac{6}{7}\right)\cdot\frac{1}{2}\right] = \frac{1}{2},$$

$$\alpha_{13} = -(b_{12}\alpha_{23} + b_{13}\alpha_{33} + b_{14}\alpha_{43})$$

$$= -\left[\left(-\frac{2}{3}\right)\cdot\frac{1}{2} + \frac{2}{3}\cdot\frac{1}{2} + 0\cdot\frac{1}{2}\right] = 0.$$

V. $$\alpha_{22} = \frac{1}{c_{22}}\,[1 - (\alpha_{23}c_{32} + \alpha_{24}c_{42})]$$

$$= \frac{1 - \left[(1/2)\cdot 1 + \left(-\frac{1}{18}\right)\cdot(8/3)\right]}{7/3} = \frac{5}{18},$$

$$\alpha_{21} = -\frac{\alpha_{22}c_{21} + \alpha_{23}c_{31} + \alpha_{24}c_{41}}{c_{11}}$$

$$= -\frac{(5/18)\cdot 2 + (1/2)\cdot 3 + (-1/18)\cdot 1}{3} = -\frac{12}{18}.$$

VI. $\alpha_{12} = -(b_{12}\alpha_{22} + b_{13}\alpha_{32} + b_{14}\alpha_{42})$

$$= -\left[\left(-\frac{2}{3}\right)\cdot\frac{5}{18} + \frac{2}{3}\cdot\frac{11}{18} + 0\cdot\left(-\frac{5}{18}\right)\right] = -\frac{4}{18}.$$

VII. $\alpha_{11} = \frac{1}{c_{11}}\,[1 - (\alpha_{12}c_{21} + \alpha_{13}c_{31} + \alpha_{14}c_{41})]$

$$= \frac{1}{3}\left[1 + \left(-\frac{4}{18}\right)\cdot 2 + 0\cdot 3 + \frac{8}{18}\cdot 1\right] = \frac{1}{3}.$$

Answer: $A^{-1} = \frac{1}{18}\begin{bmatrix} 6 & -4 & 0 & 8 \\ -12 & 5 & 9 & -1 \\ -12 & 11 & 9 & -13 \\ -6 & -5 & 9 & 1 \end{bmatrix}$. ▲

2.8. Step Matrices and Operations Involving Them

When calculating higher-order matrices, it is expedient to partition them into blocks by means of horizontal and vertical lines. Thus we divide every matrix into matrices of lower orders which are much simpler to calculate.

A matrix partitioned into blocks is called a *step* matrix or a *hypermatrix.*

For example, (1) matrix A is partitioned into four blocks

$$A = \left[\begin{array}{cc|cc} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ \hline a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{array}\right] = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right],$$

which are square matrices

$$A_{11} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},\quad A_{12} = \begin{bmatrix} a_{13} & a_{14} \\ a_{23} & a_{24} \end{bmatrix},$$

$$A_{21} = \begin{bmatrix} a_{31} & a_{32} \\ a_{41} & a_{42} \end{bmatrix},\quad A_{22} = \begin{bmatrix} a_{33} & a_{34} \\ a_{43} & a_{44} \end{bmatrix},$$

(2) the nth-order matrix C is partitioned into four blocks

$$C = \left[\begin{array}{c|c} C_{n-1} & \mathbf{u}_{n-1} \\ \hline \mathbf{v}_{n-1} & c_{nn} \end{array}\right],$$

where C_{n-1} is a square matrix of order $n-1$, $\mathbf{u}_{n-1}$ is a vector column of order $n-1$, $\mathbf{v}_{n-1}$ is a vector row of order $n-1$, and c_{nn} is a scalar.

Such a partitioning into blocks is known as *bordering* of a matrix and a matrix is said to be *bordered.*

We can add and multiply step matrices manipulating the blocks of a step matrix as the elements of an ordinary matrix.

Assume that

$$A=\left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right] \text{ and } B=\left[\begin{array}{c|c} B_{11} & B_{12} \\ \hline B_{21} & B_{22} \end{array}\right]$$

are step matrices of the same order and partitioning. Then

$$A+B=\left[\begin{array}{c|c} A_{11}+B_{11} & A_{12}+B_{12} \\ \hline A_{21}+B_{21} & A_{22}+B_{22} \end{array}\right],$$

$$AB=\left[\begin{array}{c|c} A_{11}B_{11}+A_{12}B_{21} & A_{11}B_{12}+A_{12}B_{22} \\ \hline A_{21}B_{11}+A_{22}B_{21} & A_{21}B_{12}+A_{22}B_{22} \end{array}\right].$$

Example 1 Partition the matrices

$$A=\begin{bmatrix} 5 & 7 & -3 & -4 \\ 7 & 6 & -4 & -5 \\ 6 & 4 & -3 & -2 \\ 8 & 5 & -6 & -1 \end{bmatrix} \text{ and } B=\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \\ 1 & 3 & 5 & 7 \\ 2 & 4 & 6 & 8 \end{bmatrix}$$

into square blocks and calculate $A+B$ and AB.

△ (1) We partition the matrices A and B into square blocks as follows:

$$A=\left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right]=\left[\begin{array}{cc|cc} 5 & 7 & -3 & -4 \\ 7 & 6 & -4 & -5 \\ \hline 6 & 4 & -3 & -2 \\ 8 & 5 & -6 & -1 \end{array}\right],$$

$$B=\left[\begin{array}{c|c} B_{11} & B_{12} \\ \hline B_{21} & B_{22} \end{array}\right]=\left[\begin{array}{cc|cc} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \\ \hline 1 & 3 & 5 & 7 \\ 2 & 4 & 6 & 8 \end{array}\right].$$

(2) We find that

$$A+B=\left[\begin{array}{c|c} A_{11}+B_{11} & A_{12}+B_{12} \\ \hline A_{21}+B_{21} & A_{22}+B_{22} \end{array}\right]=\left[\begin{array}{cc|cc} 6 & 9 & 0 & 0 \\ 9 & 9 & 0 & 0 \\ \hline 7 & 7 & 2 & 5 \\ 10 & 9 & 0 & 7 \end{array}\right].$$

(3) We have

$$AB=\left[\begin{array}{c|c} A_{11}B_{11}+A_{12}B_{21} & A_{11}B_{12}+A_{12}B_{22} \\ \hline A_{21}B_{11}+A_{22}B_{21} & A_{21}B_{12}+A_{22}B_{22} \end{array}\right].$$

We find in succession that

$$A_{11}B_{11}=\begin{bmatrix}5&7\\7&6\end{bmatrix}\begin{bmatrix}1&2\\2&3\end{bmatrix}=\begin{bmatrix}19&31\\19&32\end{bmatrix},$$

$$A_{12}B_{21}=\begin{bmatrix}-3&-4\\-4&-5\end{bmatrix}\begin{bmatrix}1&3\\2&4\end{bmatrix}=\begin{bmatrix}-11&-25\\-14&-32\end{bmatrix},$$

$$A_{11}B_{11}+A_{12}B_{21}=\begin{bmatrix}8&6\\5&0\end{bmatrix},$$

$$A_{21}B_{11}=\begin{bmatrix}6&4\\8&5\end{bmatrix}\begin{bmatrix}1&2\\2&3\end{bmatrix}=\begin{bmatrix}14&24\\18&31\end{bmatrix},$$

$$A_{22}B_{21}=\begin{bmatrix}-3&-2\\-6&-1\end{bmatrix}\begin{bmatrix}1&3\\2&4\end{bmatrix}=\begin{bmatrix}-7&-17\\-8&-22\end{bmatrix},$$

$$A_{21}B_{11}+A_{22}B_{21}=\begin{bmatrix}7&7\\10&9\end{bmatrix},$$

$$A_{11}B_{12}=\begin{bmatrix}5&7\\7&6\end{bmatrix}\begin{bmatrix}3&4\\4&5\end{bmatrix}=\begin{bmatrix}43&55\\45&58\end{bmatrix},$$

$$A_{12}B_{22}=\begin{bmatrix}-3&-4\\-4&-5\end{bmatrix}\begin{bmatrix}5&7\\6&8\end{bmatrix}=\begin{bmatrix}-39&-53\\-50&-68\end{bmatrix},$$

$$A_{11}B_{12}+A_{12}B_{22}=\begin{bmatrix}4&2\\-5&-10\end{bmatrix};$$

$$A_{21}B_{12}=\begin{bmatrix}6&4\\8&5\end{bmatrix}\begin{bmatrix}3&4\\4&5\end{bmatrix}=\begin{bmatrix}34&44\\44&57\end{bmatrix},$$

$$A_{22}B_{22}=\begin{bmatrix}-3&-2\\-6&-1\end{bmatrix}\begin{bmatrix}5&7\\6&8\end{bmatrix}=\begin{bmatrix}-27&-37\\-36&-50\end{bmatrix},$$

$$A_{21}B_{12}+A_{22}B_{22}=\begin{bmatrix}7&7\\8&7\end{bmatrix}.$$

Thus

$$AB=\left[\begin{array}{cc|cc}8&6&4&2\\5&0&-5&-10\\\hline 7&7&7&7\\10&9&8&7\end{array}\right].\ \blacktriangle$$

Assume that

$$A=\left[\begin{array}{c|c}A_{n-1}&\mathbf{u}_{n-1}\\\hline \mathbf{v}_{n-1}&a_{nn}\end{array}\right]\text{ and }B=\left[\begin{array}{c|c}B_{n-1}&\mathbf{y}_{n-1}\\\hline \mathbf{x}_{n-1}&b_{nn}\end{array}\right],$$

are bordered step matrices of order n and of the same partitioning. Then

$$A+B=\left[\begin{array}{c|c}A_{n-1}+B_{n-1}&\mathbf{u}_{n-1}+\mathbf{y}_{n-1}\\\hline \mathbf{v}_{n-1}+\mathbf{x}_{n-1}&a_{nn}+b_{nn}\end{array}\right],$$

where $A_{n-1} + B_{n-1}$ is a square matrix of order $n-1$,
$\mathbf{u}_{n-1} + \mathbf{y}_{n-1}$ is a vector column of order $n-1$,
$\mathbf{v}_{n-1} + \mathbf{x}_{n-1}$ is a vector row of order $n-1$,
$a_{nn} + b_{nn}$ are scalars;

$$AB = \left[\begin{array}{c|c} A_{n-1}B_{n-1}+\mathbf{u}_{n-1}\mathbf{x}_{n-1} & A_{n-1}\mathbf{y}_{n-1}+\mathbf{u}_{n-1}b_{nn} \\ \hline \mathbf{v}_{n-1}B_{n-1}+a_{nn}\mathbf{x}_{n-1} & \mathbf{v}_{n-1}\mathbf{y}_{n-1}+a_{nn}b_{nn} \end{array}\right],$$

where $A_{n-1}B_{n-1} + \mathbf{u}_{n-1}\mathbf{x}_{n-1}$ is a square matrix of order $n-1$,

$A_{n-1}\mathbf{y}_{n-1} + \mathbf{u}_{n-1}b_{nn}$ is a vector column of order $n-1$,
$\mathbf{v}_{n-1}B_{n-1} + a_{nn}\mathbf{x}_{n-1}$ is a vector row of order $n-1$,
$\mathbf{v}_{n-1}\mathbf{y}_{n-1} + a_{nn}b_{nn}$ are scalars.

Example 2. Partition the matrices

$$A = \begin{bmatrix} 5 & 8 & -4 \\ 6 & 9 & -5 \\ 4 & 7 & -3 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 3 & 2 & 5 \\ 4 & -1 & 3 \\ 9 & 6 & 5 \end{bmatrix}$$

into blocks by means of bordering and calculate $A + B$ and AB.

△ (1) We partition the matrices A and B by means of bordering:

$$A = \left[\begin{array}{cc|c} 5 & 8 & -4 \\ 6 & 9 & -5 \\ \hline 4 & 7 & -3 \end{array}\right] = \left[\begin{array}{c|c} A_2 & \mathbf{u}_2 \\ \hline \mathbf{v}_2 & a_{33} \end{array}\right], \quad B = \left[\begin{array}{cc|c} 3 & 2 & 5 \\ 4 & -1 & 3 \\ \hline 9 & 6 & 5 \end{array}\right] = \left[\begin{array}{c|c} B_2 & \mathbf{y}_2 \\ \hline \mathbf{x}_2 & b_{33} \end{array}\right].$$

(2) We find that

$$A + B = \begin{bmatrix} A_2 + B_2 & \mathbf{u}_2 + \mathbf{y}_2 \\ \mathbf{v}_2 + \mathbf{x}_2 & a_{33} + b_{33} \end{bmatrix} = \left[\begin{array}{cc|c} 8 & 10 & 1 \\ 10 & 8 & -2 \\ \hline 13 & 13 & 2 \end{array}\right].$$

(3) We have

$$AB = \begin{bmatrix} A_2B_2 + \mathbf{u}_2\mathbf{x}_2 & A_2\mathbf{y}_2 + \mathbf{u}_2b_{33} \\ \mathbf{v}_2B_2 + a_{33}\mathbf{x}_2 & \mathbf{v}_2\mathbf{y}_2 + a_{33}b_{33} \end{bmatrix}.$$

We find in succession that

$$A_2B_2 = \begin{bmatrix} 5 & 8 \\ 6 & 9 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ 4 & -1 \end{bmatrix} = \begin{bmatrix} 47 & 2 \\ 54 & 3 \end{bmatrix},$$

$$\mathbf{u}_2\mathbf{x}_2 = \begin{bmatrix} -4 \\ -5 \end{bmatrix} [9 \;\; 6] = \begin{bmatrix} -36 & -24 \\ -45 & -30 \end{bmatrix},$$

$$A_2B_2 + \mathbf{u}_2\mathbf{x}_2 = \begin{bmatrix} 11 & -22 \\ 9 & -27 \end{bmatrix},$$

$$A_2\mathbf{y}_2 = \begin{bmatrix} 5 & 8 \\ 6 & 9 \end{bmatrix} \begin{bmatrix} 5 \\ 3 \end{bmatrix} = \begin{bmatrix} 49 \\ 57 \end{bmatrix},$$

$$\mathbf{u}_2b_{33} = \begin{bmatrix} -4 \\ -5 \end{bmatrix} 5 = \begin{bmatrix} -20 \\ -25 \end{bmatrix}, \qquad A_2\mathbf{y}_2 + \mathbf{u}_2b_{33} = \begin{bmatrix} 29 \\ 32 \end{bmatrix},$$

$$\mathbf{v}_2B_2=[4\ \ 7]\begin{bmatrix}3 & 2\\ 4 & -1\end{bmatrix}=[40\ \ 1],\quad a_{33}\mathbf{x}_2=(-3)\cdot[9\ \ 6]=[-27\ \ -18],$$

$$\mathbf{v}_2B_2+a_{33}\mathbf{x}_2=[13\ \ -17],$$

$$\mathbf{v}_2\mathbf{y}_2=[4\ \ 7]\begin{bmatrix}5\\ 3\end{bmatrix}=41,\quad a_{33}b_{33}=(-3)\cdot5=-15,$$

$$\mathbf{v}_2\mathbf{y}_2+a_{33}b_{33}=26.$$

Thus

$$\left[\begin{array}{cc|c}11 & -22 & 29\\ 9 & -27 & 32\\ \hline 13 & -17 & 26\end{array}\right].\ \blacktriangle$$

2.9. Inversion of a Matrix by Partitioning It into Blocks

To find an inverse matrix, we can use the **method of partitioning into blocks**. Let $A=\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}$ is a nonsingular step matrix of order n in which A_{11} and A_{22} are square blocks of orders p and q (where $p+q=n$). We have to find the inverse matrix $A^{-1}=\begin{bmatrix}B_{11} & B_{12}\\ B_{21} & B_{22}\end{bmatrix}$, in which B_{11} and B_{22} are square matrices of orders p and q too.

According to the definition of the inverse matrix, $AA^{-1}=I_n$. In this case, the identity matrix will also be partitioned into blocks in the same way, i.e. $I_n=\begin{bmatrix}I_p & 0\\ 0 & I_q\end{bmatrix}$, where I_p and I_q are identity matrices of orders p and q respectively. Then

$$A^{-1}A=\begin{bmatrix}B_{11} & B_{12}\\ B_{21} & B_{22}\end{bmatrix}\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}=\begin{bmatrix}I_p & 0\\ 0 & I_q\end{bmatrix},$$

whence, after multiplication, we get four matrix equations

$$\begin{cases}B_{11}A_{11}+B_{12}A_{21}=I_p,\\ B_{11}A_{12}+B_{12}A_{22}=0,\\ B_{21}A_{11}+B_{22}A_{21}=0,\\ B_{21}A_{12}+B_{22}A_{22}=I_q.\end{cases}\qquad(1)$$

To find the blocks of the matrix A^{-1}, we have to solve the system of matrix equations (1). To do this, we use the method of elimination of the unknowns. We postmultiply the first equation of system (1) by $A_{11}^{-1}A_{12}$ and subtract the second equation of the system from the result of the multiplication. We obtain

$$B_{12}(A_{21}A_{11}^{-1}A_{12} - A_{22}) = A_{11}^{-1}A_{12}.$$

From this we find that

$$B_{12} = -A_{11}^{-1}A_{12}(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1},$$
$$B_{11} = A_{11}^{-1} - B_{12}A_{21}A_{11}^{-1}.$$

Similarly, we find from the third and fourth equations of the system that

$$B_{22} = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1},$$
$$B_{21} = -B_{22}A_{21}A_{11}^{-1}.$$

This is possible under the condition that the requisite operations have sense. We introduce the following designations:

$$X = A_{11}^{-1}A_{12}, \quad Y = A_{21}A_{11}^{-1},$$
$$\Theta = A_{22} - A_{21}X = A_{22} - YA_{12}.$$

Then the formulas for the blocks of the inverse matrix A^{-1} can be written in the following form:

$$B_{11} = A_{11}^{-1} + X\Theta^{-1}Y, \quad B_{12} = -X\Theta^{-1},$$
$$B_{21} = -\Theta^{-1}Y, \quad B_{22} = \Theta^{-1}. \qquad (2)$$

Formulas (2) are valid provided that A_{11}^{-1} and Θ^{-1} exist. It is convenient to make the following scheme for calculations:

	A_{21}	A_{22}
$X = A_{11}^{-1}A_{12}$	A_{11}^{-1}	A_{12}
Θ^{-1}	$Y = A_{21}A_{11}^{-1}$	$\Theta = A_{22} - YA_{12}$

The inverse matrix has the form

$$A^{-1} = \left[\begin{array}{c|c} A_{11}^{-1} + X\Theta^{-1}Y & -X\Theta^{-1} \\ \hline -\Theta^{-1}Y & \Theta^{-1} \end{array}\right].$$

Example 1. By means of partitioning into blocks, invert the matrix

$$A=\left[\begin{array}{rr|rr} 1 & 0 & 1 & 2 \\ -1 & 2 & 3 & 1 \\ \hline 4 & 0 & -2 & 1 \\ 0 & 2 & 1 & 2 \end{array}\right].$$

△ We designate

$$A_{11}=\begin{bmatrix} 1 & 0 \\ -1 & 2 \end{bmatrix}, \quad A_{12}=\begin{bmatrix} 1 & 2 \\ 3 & 1 \end{bmatrix}, \quad A_{21}=\begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}, \quad A_{22}=\begin{bmatrix} -2 & 1 \\ 1 & 2 \end{bmatrix}$$

and make the necessary calculations:

$$\det A_{11}=2, \quad A_{11}^{-1}=\frac{1}{2}\begin{bmatrix} 2 & 0 \\ 1 & 1 \end{bmatrix},$$

$$X=A_{11}^{-1}A_{12}=\frac{1}{2}\begin{bmatrix} 2 & 0 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 3 & 1 \end{bmatrix}=\frac{1}{2}\begin{bmatrix} 2 & 4 \\ 4 & 3 \end{bmatrix},$$

$$Y=A_{21}A_{11}^{-1}=\frac{1}{2}\begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} 2 & 0 \\ 1 & 1 \end{bmatrix}=\begin{bmatrix} 4 & 0 \\ 1 & 1 \end{bmatrix},$$

$$\Theta=A_{22}-YA_{12}=\begin{bmatrix} -2 & 1 \\ 1 & 2 \end{bmatrix}-\begin{bmatrix} 4 & 0 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 3 & 1 \end{bmatrix}=\begin{bmatrix} -6 & -7 \\ -3 & -1 \end{bmatrix},$$

$$\det\Theta=-15; \quad \Theta^{-1}=\frac{1}{15}\begin{bmatrix} 1 & -7 \\ -3 & 6 \end{bmatrix}.$$

We write the initial data and the results of the calculations in the form of the following table:

	$A_{21}=\begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}$	$A_{22}=\begin{bmatrix} -2 & 1 \\ 1 & 2 \end{bmatrix}$
$X=\frac{1}{2}\begin{bmatrix} 2 & 4 \\ 4 & 3 \end{bmatrix}$	$A_{11}^{-1}=\frac{1}{2}\begin{bmatrix} 2 & 0 \\ 1 & 1 \end{bmatrix}$	$A_{12}=\begin{bmatrix} 1 & 2 \\ 3 & 1 \end{bmatrix}$
$\Theta^{-1}=\frac{1}{15}\begin{bmatrix} 1 & -7 \\ -3 & 6 \end{bmatrix}$	$Y=\begin{bmatrix} 4 & 0 \\ 1 & 1 \end{bmatrix}$	$\Theta=\begin{bmatrix} -6 & -7 \\ -3 & -1 \end{bmatrix}$

Then we have

$$X\Theta^{-1}=\frac{1}{2}\begin{bmatrix} 2 & 4 \\ 4 & 3 \end{bmatrix}\cdot\frac{1}{15}\begin{bmatrix} 1 & -7 \\ -3 & 6 \end{bmatrix}=\frac{1}{30}\begin{bmatrix} -10 & 10 \\ -5 & -10 \end{bmatrix},$$

$$X\Theta^{-1}Y=\frac{1}{30}\begin{bmatrix} -10 & -10 \\ -5 & -10 \end{bmatrix}\begin{bmatrix} 4 & 0 \\ 1 & 1 \end{bmatrix}=\frac{1}{30}\begin{bmatrix} -30 & 10 \\ -30 & -10 \end{bmatrix},$$

$$\Theta^{-1}Y=\frac{1}{15}\begin{bmatrix}1 & -7\\ -3 & 6\end{bmatrix}\begin{bmatrix}4 & 0\\ 1 & 1\end{bmatrix}=\frac{1}{15}\begin{bmatrix}-3 & -7\\ -6 & 6\end{bmatrix},$$

$$X\Theta^{-1}Y=\frac{1}{2}\begin{bmatrix}2 & 4\\ 4 & 3\end{bmatrix}\cdot\frac{1}{15}\begin{bmatrix}-3 & -7\\ -6 & 6\end{bmatrix}=\frac{1}{30}\begin{bmatrix}-30 & 10\\ -30 & -10\end{bmatrix}.$$

To ensure the correctness of the result, we have used two techniques to calculate the product $X\Theta^{-1}Y$: $X\Theta^{-1}Y=(X\Theta^{-1})\,Y$ and $X\Theta^{-1}Y=X\,(\Theta^{-1}Y)$.

The final result is

$$A^{-1}=\begin{bmatrix}A_{11}^{-1}+X\Theta^{-1}Y & -X\Theta^{-1}\\ -\Theta^{-1}Y & \Theta^{-1}\end{bmatrix}$$

$$=\begin{bmatrix}\frac{1}{30}\begin{bmatrix}0 & 10\\ -15 & 5\end{bmatrix} & \frac{1}{30}\begin{bmatrix}10 & -10\\ 5 & 10\end{bmatrix}\\ \frac{1}{15}\begin{bmatrix}3 & 7\\ -6 & 6\end{bmatrix} & \frac{1}{15}\begin{bmatrix}1 & -7\\ -3 & 6\end{bmatrix}\end{bmatrix}$$

$$\frac{1}{30}\begin{bmatrix}0 & 10 & 10 & -10\\ -15 & 5 & 5 & 10\\ 0 & 14 & 2 & -14\\ 12 & -12 & -6 & 12\end{bmatrix}. \blacktriangle$$

The method of successive bordering is a special case of the method of inversion of step matrices.

Assume that we are given a nonsingular square matrix of order n

$$A=\left|\begin{matrix}a_{11} & a_{12} & \dots & a_{1n}\\ a_{21} & a_{22} & \dots & a_{2n}\\ \cdot & \cdot & \cdot & \cdot\\ a_{n1} & a_{n2} & \dots & a_{nn}\end{matrix}\right|,$$

for which we have to find the inverse matrix A^{-1}.

We compose the succession of matrices:

$$A_1=[a_{11}],$$

$$A_2=\begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix},$$

$$A_3=\begin{bmatrix}a_{11} & a_{12} & a_{13}\\ a_{21} & a_{22} & a_{23}\\ a_{31} & a_{32} & a_{33}\end{bmatrix}=\left[\begin{array}{cc|c}\multicolumn{2}{c|}{A_2} & \begin{matrix}a_{13}\\ a_{23}\end{matrix}\\ \hline a_{31} & a_{32} & a_{33}\end{array}\right],$$

$$A_4=\begin{bmatrix}a_{11} & a_{12} & a_{13} & a_{14}\\ a_{21} & a_{22} & a_{23} & a_{24}\\ a_{31} & a_{32} & a_{33} & a_{34}\\ a_{41} & a_{42} & a_{43} & a_{44}\end{bmatrix}=\left[\begin{array}{ccc|c}\multicolumn{3}{c|}{A_3} & \begin{matrix}a_{14}\\ a_{24}\\ a_{34}\end{matrix}\\ \hline a_{41} & a_{42} & a_{43} & a_{44}\end{array}\right],$$

$$A_2^{-1} = B_2 = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix},$$

$$A_3^{-1} = C_3 = \left[\begin{array}{cc|c} \multicolumn{2}{c|}{C_2} & c_{13} \\ & & c_{23} \\ \hline c_{31} & c_{32} & c_{33} \end{array}\right],$$

$$A_4^{-1} = D_4 = \left[\begin{array}{ccc|c} & & & d_{14} \\ \multicolumn{3}{c|}{D_3} & d_{24} \\ & & & d_{34} \\ \hline d_{41} & d_{42} & d_{43} & d_{44} \end{array}\right]$$

and so on. Each successive matrix is obtained from the preceding one by means of bordering.

The inverse of the second matrix $A_2^{-1} = B_2$ can be found in a direct way:

$$A_2^{-1} = B_2 = \begin{bmatrix} a_{22}/|A_2| & -a_{12}/|A_2| \\ -a_{21}/|A_2| & a_{11}/|A_2| \end{bmatrix},$$

where

$$|A_2| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

Using the matrix A_2^{-1} and applying to A_3 the scheme for calculation presented above, we can get A_3^{-1}, then using A_3^{-1}, we can similarly get A_4^{-1}, ... , and finally get $A_n^{-1} = A^{-1}$.

Example 2. Using the method of successive bordering, invert the matrix

$$A = \begin{bmatrix} 3 & -4 & 5 & 0 \\ 2 & -3 & 1 & 1 \\ 3 & -5 & -1 & 2 \\ 3 & -1 & 4 & 1 \end{bmatrix}.$$

△ We have

$$A = \left[\begin{array}{cc|c|c} -3 & -4 & 5 & 0 \\ 2 & -3 & 1 & 1 \\ \hline 3 & -5 & -1 & 2 \\ \hline 3 & -1 & 4 & 1 \end{array}\right].$$

(1) $A_2 = \begin{bmatrix} 3 & -4 \\ 2 & -3 \end{bmatrix}$, det $A_2 = -9+8 = -1$,

$$A_2^{-1} = B_2 = \begin{bmatrix} 3 & -4 \\ 2 & -3 \end{bmatrix}.$$

(2) The scheme for calculating $A_3^{-1} = C_3$ has the following form:

	$[a_{31} a_{32}]$	a_{33}
X	A_2^{-1}	$\begin{bmatrix} a_{13} \\ a_{23} \end{bmatrix}$
Θ^{-1}	Y	Θ

We carry out the calculations

$$X = A_2^{-1} \begin{bmatrix} a_{13} \\ a_{23} \end{bmatrix} = \begin{bmatrix} 3 & -4 \\ 2 & -3 \end{bmatrix} \begin{bmatrix} 5 \\ 1 \end{bmatrix} = \begin{bmatrix} 11 \\ 7 \end{bmatrix},$$

$$Y = [a_{31} \; a_{32}] \, A_2^{-1} = [3 \; -5] \begin{bmatrix} 3 & -4 \\ 2 & -3 \end{bmatrix} = [-1 \; 3],$$

$$\Theta = a_{33} - Y \begin{bmatrix} a_{13} \\ a_{23} \end{bmatrix} = -1 - [-1 \; 3] \begin{bmatrix} 5 \\ 1 \end{bmatrix} = -1 + 2 = 1,$$

$$\Theta^{-1} = 1, \quad X\Theta^{-1}Y = \begin{bmatrix} 11 \\ 7 \end{bmatrix} \cdot 1 \cdot [-1 \; 3] = \begin{bmatrix} -11 & 33 \\ -7 & 21 \end{bmatrix}.$$

Then we fill in the table:

	3 −5	−1
11	3 −4	5
7	2 −3	1
1	−1 3	1

We obtain the elements of the inverse matrix $A_3^{-1} = C_3$ after making the following matrix additions and multiplications:

$$c_{33} = \Theta^{-1} = 1, \quad [c_{31} \; c_{32}] = -\Theta^{-1}Y = [1 \; -3],$$

$$\begin{bmatrix} c_{13} \\ c_{23} \end{bmatrix} = -X\Theta^{-1} = \begin{bmatrix} -11 \\ -7 \end{bmatrix},$$

$$C_2 = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} = A_2^{-1} + X\Theta^{-1}Y$$

$$= \begin{bmatrix} 3 & -4 \\ 2 & -3 \end{bmatrix} + \begin{bmatrix} -11 & 33 \\ -7 & 21 \end{bmatrix} = \begin{bmatrix} -8 & 29 \\ -5 & 18 \end{bmatrix}.$$

Thus

$$A_3^{-1}=C_3=\begin{bmatrix}-8 & 29 & -11\\ -5 & 18 & -7\\ 1 & -3 & 1\end{bmatrix}.$$

(3) To calculate $A_4^{-1}=D_4=A^{-1}$, we compile a table:

	$[a_{41}a_{42}a_{43}]$	a_{44}
X	A_3^{-1}	$\begin{bmatrix}a_{14}\\ a_{24}\\ a_{34}\end{bmatrix}$
Θ^{-1}	Y	Θ

Then we carry out the calculations:

$$X=\begin{bmatrix}-8 & 29 & -11\\ -5 & 18 & -7\\ 1 & -3 & 1\end{bmatrix}\begin{bmatrix}0\\ 1\\ 2\end{bmatrix}=\begin{bmatrix}7\\ 4\\ -1\end{bmatrix},$$

$$Y=[3\ \ -1\ \ 4]\begin{bmatrix}-8 & 29 & -11\\ -5 & 18 & -7\\ 1 & -3 & 1\end{bmatrix}=[-15\ \ 57\ \ -22],$$

$$\Theta=1-[-15\ \ 57\ \ -22]\begin{bmatrix}0\\ 1\\ 2\end{bmatrix}=1-13=-12,\quad \Theta^{-1}=-1/12.$$

We fill in the table:

	3	−1	4	1
7	−8	29	11	0
4	−5	18	−7	1
−1	1	−3	1	2
−1/12	−15	57	−22	−12

Then we have

$$d_{44} = -\frac{1}{12}; \quad [d_{41} d_{42} d_{43}] = \frac{1}{12}[-15 \quad 57 \quad -22],$$

$$\begin{bmatrix} d_{14} \\ d_{24} \\ d_{34} \end{bmatrix} = \frac{1}{12}\begin{bmatrix} 7 \\ 4 \\ -1 \end{bmatrix},$$

$$D_3 = A_3^{-1} + X\Theta^{-1}Y = \begin{bmatrix} -8 & 29 & 11 \\ -5 & 18 & -7 \\ 1 & -3 & 1 \end{bmatrix} - \frac{1}{12}\begin{bmatrix} 7 \\ 4 \\ -1 \end{bmatrix}[-15 \quad 57 \quad -22]$$

$$= \frac{1}{12}\begin{bmatrix} -96 & 348 & -132 \\ -60 & 216 & -84 \\ 12 & -36 & 12 \end{bmatrix} + \frac{1}{12}\begin{bmatrix} 105 & -399 & 154 \\ 60 & -228 & 88 \\ -15 & 57 & -22 \end{bmatrix}$$

$$= \frac{1}{12}\begin{bmatrix} 9 & -51 & 22 \\ 0 & -12 & 4 \\ -3 & 21 & 10 \end{bmatrix}.$$

Thus

$$A^{-1} = \frac{1}{12}\begin{bmatrix} 9 & -51 & 22 & 7 \\ 0 & -12 & 4 & 4 \\ -3 & 21 & -10 & -1 \\ -15 & 57 & -22 & -1 \end{bmatrix}. \quad \blacktriangle$$

The method of bordering can only be used if all the intermediate matrices $A_2, A_3, \ldots, A_{n-1}$ are nonsingular.

2.10. The Absolute Value and the Norm of a Matrix

The *absolute value* (*modulus*) *of the matrix* $A = [a_{ij}]$ is a matrix $|A| = [|a_{ij}|]$, where all the elements $|a_{ij}|$ are the moduli of the elements of the matrix A.

Let A and B be matrices for which the operations $A + B$ and AB have sense. Then

(1°) $|A + B| \leqslant |A| + |B|$,

(2°) $|AB| < |A| \cdot |B|$,

(3°) $|\alpha A| = |\alpha| \cdot |A|$, where α is a scalar.

The *norm* of the matrix $A = [a_{ij}]$ is a real number $\|A\|$, which satisfies the following conditions:

(1°) $\|A\| \geqslant 0$ (with $\|A\| = 0$ if and only if $A = 0$),

(2°) $\|\alpha A\| = |\alpha| \cdot \|A\|$, where α is a scalar (with $\|-A\| = \|A\|$),

(3°) $\|A + B\| \leqslant \|A\| + \|B\|$, (4°) $\| AB \| \leqslant \| A \| \cdot \| B \|$,

(5°) $\| A - B \| \geqslant | \| B \| - \| A \| |$ where A and B are matrices for which the corresponding operations have sense.

Let us consider the following three norms, which are easy to calculate, for the matrix $A = [a_{ij}]$ of an arbitrary dimension:

$\| A \|_1 = \max\limits_i \sum\limits_j | a_{ij} |$ is the *maximum sum of the moduli of the elements of the matrix by rows*,

$\| A \|_2 = \max\limits_j \sum\limits_i | a_{ij} |$ is the *maximum sum of the moduli of the elements of the matrix by the columns*,

$\| A \|_3 = \sqrt{\sum\limits_{ij} | a_{ij} |^2}$ is the *square root of the sum of the squares of the moduli of all the elements of the matrix.*

Example 1. For the matrix

$$A = \begin{bmatrix} 2 & -1 & 4 \\ 5 & 3 & 2 \\ 6 & -7 & 3 \end{bmatrix}$$

calculate $\| A \|_1$, $\| A \|_2$ and $\| A \|_3$.

△ We find that

$$\| A \|_1 = \max(2+1+4,\ 5+3+2,\ 6+7+3) = \max(7,\ 10,\ 16) = 16,$$

$$\| A \|_2 = \max(2+5+6,\ 1+3+7,\ 4+2+3) = \max(13,\ 11,\ 9) = 13,$$

$$\| A \|_3 = \sqrt{2^2+1^2+4^2+5^2+3^2+2^2+6^2+7^2+3^2} = \sqrt{153} = 12.2. \ \blacktriangle$$

For the vector $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ these norms can be calculated by the following formulas:

$\| \mathbf{x} \|_1 = \max | x_i |$ is the *coordinate of the vector maximum in its absolute value*,

$\| \mathbf{x} \|_2 = | x_1 | + | x_2 | + \ldots + | x_n |$ is the *sum of the moduli of the coordinates of the vector*,

$\| \mathbf{x} \|_3 = \sqrt{| x_1 |^2 + | x_2 |^2 + \ldots + | x_n |^2}$ is the *square root of the sum of the squares of the moduli of the coordinates of the vector.*

The norm $\| \mathbf{x} \|_3$ is the *absolute value of the vector.*

Example 2. For the vector $\mathbf{x}=\begin{bmatrix}1\\2\\3\\-5\end{bmatrix}$ calculate $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2$ and $\|\mathbf{x}\|_3$.

△ We have

$$\|\mathbf{x}\|_1=\max(1,\ 2,\ 3,\ 5)=5,\quad \|\mathbf{x}_2\|=1+2+3+5=11,$$

$$\|\mathbf{x}\|_3=\sqrt{1^2+2^2+3^2+5^2}=\sqrt{39}=6.2.\ \blacktriangle$$

2.11. The Rank of a Matrix and the Methods of Its Calculation

Consider a rectangular matrix

$$A=\begin{bmatrix}a_{11} & a_{12} & \cdots & a_{1n}\\ a_{21} & a_{22} & \cdots & a_{2n}\\ \cdot & \cdot & \cdot & \cdot\\ a_{m1} & a_{m2} & \cdots & a_{mn}\end{bmatrix}.$$

If we choose k arbitrary rows and k arbitrary columns in this matrix, where $k \leqslant \min(m,\ n)$, then the elements, which are at the intersections of these rows and columns, form a square matrix of order k whose determinant is a kth-order *minor* of the matrix A.

For instance, the intersection of the first and second rows with the first and second columns of the matrix A is occupied by a second-order matrix $\begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}$ whose determinant is a second-order minor of the matrix A. We designate it as M_2:

$$M_2=\begin{vmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{vmatrix}.$$

The *rank* of the matrix A is the maximum order of the nonzero minor of this matrix. It follows from the definition of the rank that if the rank of a matrix is r, then there is at least one nonzero minor of order r in the matrix and all the minors of order $r+1$ and of higher orders are zero. Note that the rank of a zero matrix is zero and that of a nonzero row matrix (or column matrix) is equal to unity.

For a rectangular matrix of dimension $m \times n$ the difference of the least of the numbers m and n and the rank of the matrix is known as the *deficiency* of the matrix.

For an $n \times n$ square matrix the deficiency is equal to $n - r$. If the deficiency is zero, then the rank of the matrix is the greatest of the possible ranks for that dimension.

Let us consider one of the methods of calculating the rank of a matrix based on the definition of the rank. In this case it is necessary to find the minor of the maximum order different from zero. When using this method to calculate the rank of a matrix, we pass from minors of lower orders (beginning with the minors of the first order, i.e. the elements of the matrix) to the minors of higher orders, carrying out the operations according to the following rule: assume that we have found the rth-order minor M_r, different from zero; then we have only to calculate the minors of order $r+1$ which border the minor M_r. If all these minors are zero, then the rank of the matrix is r, but if at least one of them is nonzero, then the operation must involve that minor, and in that case the rank of the matrix is obviously higher than r. This method of calculation is known as the **method of bordering.**

Example 1. Using the method of bordering, find the rank of the matrix

$$A = \begin{bmatrix} 2 & 4 & 1 & 0 \\ 2 & 4 & 1 & 0 \\ -1 & -2 & 3 & 1 \\ 0 & 0 & 7 & 2 \end{bmatrix}$$

△ (1) We choose a first-order minor which is nonzero: $M_1^1 = a_{22} = 4 \neq 0$ (the upper index is for the ordinal number in the calculation and the lower index is for the order of the minor).

(2 We find the second-order bordering minor which is nonzero:

$$M_2^1 = \begin{bmatrix} 4 & 1 \\ -2 & 3 \end{bmatrix} \neq 0.$$

(3) We consider all third-order minors which border the minor M_2^1, for which purpose we compose minors M_3^1 and M_3^2 from the second, the third and the fourth row:

$$M_3^1 = \begin{vmatrix} 2 & 4 & 1 \\ -1 & -2 & 3 \\ 0 & 0 & 7 \end{vmatrix} = 0, \quad M_3^2 = \begin{vmatrix} 4 & 1 & 0 \\ -2 & 3 & 1 \\ 0 & 7 & 2 \end{vmatrix} = 0$$

since in these minors the third row is equal to the sum of the first row and the doubled second row.

The bordering minors from the first, the second and the third row are also equal to zero since the first and the second rows are identical. If all the third-order minors are zero, then all the higher-order minors (beginning with $3+1=4$) are also zero. Consequently, the rank of the matrix is 2 and the deficiency $4-2=2$. ▲

Since the number of determinants of different orders, generated by the matrix, is usually large, the calculation of the rank by the method of bordering is very laborious.

It is much simpler to calculate the rank by means of *elementary transformations of the matrix.*

The elementary transformations are

(1) permutation of two rows (columns),

(2) multiplication of a row (column) by a number k ($k \neq 0$),

(3) addition of a row, multiplied by a nonzero number k, to another row (column),

(4) elimination of a row (column), consisting of zeros, from the matrix,

(5) elimination of a row (column) which is a linear combination of other rows (columns) from a matrix.

Elementary transformations do not change the rank of a matrix, i.e. transformations of this kind result in a new matrix which is not equal to the original one but is equivalent to it (the ranks of these matrices are equal). The symbol $\sim$ is used to denote the equivalence of matrices.

Transformations of a matrix can be carried out in the following order.

1. If the element a_{11} in the upper left corner of the matrix is zero and the matrix has nonzero elements, then by interchanging rows and columns we replace the zero element by one of the nonzero elements and, using elementary transformations, turn all the other elements of the first row and all the elements of the first column into zeros. In what follows, the first column and the first row remain unchanged (we can only interchange them). If all the other elements of the transformed matrix are zero, the rank of the matrix is equal to unity, i.e. $r=1$.

2. If $a_{22}=0$ in the transformed matrix but there are nonzero elements, we can put one of them at the intersection of the second row and the second column and

apply elementary transformations to turn into zeros first all the other elements of the second row and then all the other elements of the second column and continue with the transformations in a similar way.

As a result we can obtain a matrix whose elements are only unities which are on the principal diagonal, the number of these elements being equal to the rank of the matrix, and whose other elements are zeros.

Example 2. Using elementary transformations, determine the rank of the matrix

$$A=\begin{bmatrix} 1 & 2 & 3 & -1 \\ -1 & 3 & 2 & 0 \\ 3 & 1 & 4 & -2 \\ -3 & 4 & 1 & 1 \end{bmatrix}.$$

△ Let us carry out in succession the following elementary transformations of the matrix.

(a) First we add the first column to the fourth and then, multiplying in succession by (—2) and (—3), add the results to the second and the third column respectively.

(b) We eliminate the second and the third column since they result from the fourth column after the multiplication by (—5).

Then we have

$$A=\begin{bmatrix} 1 & 2 & 3 & -1 \\ -1 & 3 & 2 & 0 \\ 3 & 1 & 4 & -2 \\ -3 & 4 & 1 & 1 \end{bmatrix} \overset{(a)}{\sim} \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 5 & 5 & -1 \\ 3 & -5 & -5 & 1 \\ -3 & 10 & 10 & -2 \end{bmatrix} \overset{(b)}{\sim} \begin{bmatrix} 1 & 0 \\ -1 & -1 \\ 3 & 1 \\ -3 & -2 \end{bmatrix}.$$

It is evident that the rank of the last matrix is 2. It cannot be higher than 2 since if $r = 2$, then the deficiency is zero ($n - r = 0$) and cannot be lower than 2 since the matrix has a minor $M_2 \neq 0$.

The matrix obtained is equivalent to the matrix A and, consequently, $r(A) = 2$.

We can arrive at the same result if we continue with the transformations of the matrix.

(c) First we add the first row to the second and then, multiplying it in succession by (—3) and 3, add the results to the third and the fourth row respectively.

(d) First we add the third row to the second and then, multiplying it by 2, add the result to the fourth row.

(e) We eliminate the zero rows.

We have

$$A \sim \begin{bmatrix} 1 & 0 \\ -1 & -1 \\ 3 & 1 \\ -3 & -2 \end{bmatrix} \overset{(c)}{\sim} \begin{bmatrix} 1 & 0 \\ 0 & -1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \overset{(d)}{\sim} \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \overset{(e)}{\sim} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The last matrix does not play the part of an identity matrix, it results from elementary transformations and is equivalent to the matrix A. The number of unities on the principal diagonal of the matrix obtained is 2. Consequently, $r(A) = 2$. ▲

2.12. The Concept of a Linear (Vector) Space. The Linear Dependence of Vectors

The *linear* (*vector*) *space* is the set U of elements $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$, ... for which the operations of addition of the elements and the multiplication of them by a real number can be performed within the set U and which satisfy the following axioms:

(1°) $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$,

(2°) $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$,

(3°) there is an element $\mathbf{0} \in U$ such that $\mathbf{x} + \mathbf{0} = \mathbf{x}$,

(4°) for every $\mathbf{x}$ there is an opposite element $-\mathbf{x} \in U$ such that $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$,

(5°) $1 \cdot \mathbf{x} = \mathbf{x}$,

(6°) $\alpha(\beta\mathbf{x}) = (\alpha\beta)\mathbf{x}$,

(7°) $(\alpha + \beta)\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{x}$,

(8°) $\alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y}$, where $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z} \in U$ and α and β are real numbers.

Note that $0 \cdot \mathbf{x} = \mathbf{0}$ and $(-1)\mathbf{x} = -\mathbf{x}$.

Example 1. The set of all n-dimensional vectors with the ordinary operations of addition of vectors and multiplication of a vector by a real number α forms a linear space since these operations satisfy axioms 1°-8°.

The sum of two vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ is a vector $\mathbf{z} = \mathbf{x} + \mathbf{y} = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)$; the product of the vector $\mathbf{x}$ by the number α is a vector $\alpha\mathbf{x} = (\alpha x_1, \alpha x_2, \ldots, \alpha x_n)$ (see 2.1). The vector $\mathbf{0} = (0, 0, \ldots, 0)$ is a null vector; the vector $-\mathbf{x} = (-x_1, -x_2, \ldots, -x_n)$ is opposite to the vector $\mathbf{x}$. All the axioms of the linear space are evidently satisfied.

Example 2. The set of vectors of different dimensions is not a linear space since the operation of addition is not defined for them.

Example 3. It can be shown that square matrices of the same order form a linear space whereas square matrices of different orders do not form a linear space since the sum is not defined for them.

The notion of linear dependence of vectors is of great importance in linear (vector) space.

If $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are vectors from the space U, then the vector $\mathbf{z} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \ldots + c_n\mathbf{x}_n = \sum\limits_{i=1}^{n} c_i\mathbf{x}_i$ is a *linear combination* of the vectors $\mathbf{x}_i$.

If the vector $\mathbf{z} = \sum\limits_{i=1}^{n} c_i\mathbf{x}_i = 0$ but among the numbers $c_1, c_2, \ldots, c_n$ there is at least one number different from zero, then the vectors $\mathbf{x}_i$ are said to be *linearly dependent.*

The vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are *linearly independent* if their linear combination $\mathbf{z} = \sum\limits_{i=1}^{n} c_i\mathbf{x}_i$ is equal to the zero vector if and only if all $c_i = 0$ $(i = 1, 2, \ldots, n)$.

There are not more than two linearly independent vectors on the plane and not more than three linearly independent vectors in the three-dimensional space.

Theorem. *If the vectors* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, *which belong to the linear space* U, *are linearly dependent, then at least one of them is a linear combination of the other ones.*

☐ Since the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are linearly dependent, it follows that $c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \ldots + c_n\mathbf{x}_n =$, and at least one of the numbers c_i $(i = 1, 2, \ldots, n)$ is nonzero. Let $c_k \neq 0$ $(1 \leqslant k \leqslant n)$. Then the vector $\mathbf{x}_k$ is a linear combination of the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{k-1}, \mathbf{x}_{k+1}, \ldots, \mathbf{x}_n$:

$$\mathbf{x}_k = -\frac{c_1}{c_k}\mathbf{x}_1 - \frac{c_2}{c_k}\mathbf{x}_2 - \ldots - \frac{c_{k-1}}{c_k}\mathbf{x}_{k-1} - \frac{c_{k+1}}{c_k}\mathbf{x}_{k+1} - \ldots - \frac{c_n}{c_k}\mathbf{x}_n. \quad ■$$

Example 4. Given a system of n n-dimensional unit vectors

$$\begin{aligned} \mathbf{e}_1 &= (1, 0, 0, \ldots, 0), \\ \mathbf{e}_2 &= (0, 1, 0, \ldots, 0), \\ \mathbf{e}_n &= (0, 0, 0, \ldots, 1) \end{aligned}$$

find out whether the system is linearly dependent.

△ (1) We form a linear combination of these vectors and equate it to zero:

$$c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + \ldots + c_n\mathbf{e}_n = 0.$$

(2) We write this relation in coordinates:

$$\begin{aligned} c_1\cdot 1 + c_2\cdot 0 + \ldots + c_n\cdot 0 &= 0, \\ c_1\cdot 0 + c_2\cdot 1 + \ldots + c_n\cdot 0 &= 0, \\ \\ c_1\cdot 0 + c_2\cdot 0 + \ldots + c_n\cdot 1 &= 0. \end{aligned}$$

Hence $c_1 = c_2 = c_3 = \ldots = c_n = 0$. This means that a system of n n-dimensional unit vectors is linearly independent. It is evident that any part of this system of vectors is linearly independent either. ▲

Example 5. Given vectors $\mathbf{x}_1 = (1, 2, 3)$ and $\mathbf{x}_2 = (3, 6, 7)$, find out whether these vectors are linearly dependent.

△ (1) We form a linear combination of the vectors and equate it to zero:

$$c_1\mathbf{x}_1 + c_2\mathbf{x}_2 = 0. \qquad (*)$$

(2) The vector equality $(*)$ is equivalent to the homogeneous system of three linear equations with two unknowns

$$\begin{cases} 1\cdot c_1+3c_2=0, \\ 2c_1+6c_2=0, \\ 3c_1+7c_2=0. \end{cases} \qquad (**)$$

(3) We divide both sides of the second equation of system $(**)$ by 2:

$$\begin{cases} 1\cdot c_1+3c_2=0, \\ 1\cdot c_1+3c_2=0, \\ 3c_1+7c_2=0. \end{cases} \qquad (***)$$

There are two identical equations in system $(***)$ one of which we delete.

(4) Solving now the system

$$\begin{cases} 1\cdot c_1+3c_2=0, \\ 3c_1+7c_2=0, \end{cases}$$

we get $c_1 = 0$ and $c_2 = 0$. Thus the given system of vectors is linearly independent. ▲

Example 6. Given vectors $\mathbf{x}_1 = (1, 3)$, $\mathbf{x}_2 = (0, 2)$, $\mathbf{x}_3 = (5, 7)$, find out whether this system of vectors is linearly dependent.

△ (1) We form a linear combination of the vectors and equate it to zero:

$$c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + c_3\mathbf{x}_3 = 0. \qquad (*)$$

We write relation $(*)$ in coordinates:

$$\begin{cases} 1\cdot c_1+0\cdot c_2+5c_3=0, \\ 3c_1+2c_2+7c_3=0. \end{cases} \qquad (**)$$

We have obtained a system of two linear equations with three unknowns.

(2) We use the method of substitution to solve this system. We substitute the expression $c_1 = -5c_3$, obtained from the first equation, into the second equation: $-15c_3 + 2c_2 + 7c_3 = 0$. Hence $c_2 = 4c_3$. We assign an arbitrary nonzero numerical value to c_3, say, $c_3 = 1$. Then $c_1 = -5$ and $c_2 = 4$.

(3) We substitute these values into relation $(*)$:

$$-5\mathbf{x}_1 + 4\mathbf{x}_2 + \mathbf{x}_3 = 0, \quad \text{or} \quad \mathbf{x}_3 = 5\mathbf{x}_1 - 4\mathbf{x}_2,$$

i.e. the vector $\mathbf{x}_3$ is a linear combination of the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, and this means that the vectors $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ are linearly dependent. ▲

2.13. The Basis of Space

The linear (vector) space U is said to be *n-dimensional* if there are n linearly independent vectors in it and there are no more linearly independent vectors. The number n is the *dimension of the space* and the space itself is *finite-dimensional* (designated as U_n).

If there is an arbitrary number of linearly independent vectors in the space U, then the space is *infinite-dimensional.*

Any set of n linearly independent vectors of n-dimensional space is the *basis* of that space.

In every space there is an infinite set of bases. One of them is a system of *unit vectors*:

$$\begin{aligned} \mathbf{e}_1 &= (1, 0, 0, \ldots, 0), \\ \mathbf{e}_2 &= (0, 1, 0, \ldots, 0), \\ &\cdots\cdots\cdots\cdots \\ \mathbf{e}_n &= (0, 0, 0, \ldots, 1). \end{aligned}$$

Theorem 1. *Every vector of an n-dimensional space U_n can be represented as a linear combination of the vectors of the basis, and that representation is unique.*

☐ Let $\mathbf{x} \in U_n$ and $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ be the basis of the space U_n. The vectors $\mathbf{x}, \mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ are linearly dependent (their number is $n + 1$ and exceeds the dimension of the space), i.e.

$$c_0\mathbf{x} + c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + \ldots + c_n\mathbf{e}_n = 0, \tag{1}$$

where a certain coefficient $c_j \neq 0$ $(0 \leqslant j \leqslant n)$. In relation (1) the coefficient $c_0 \neq 0$ since otherwise

$$c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + \ldots + c_n\mathbf{e}_n = 0,$$

where $c_j \neq 0$ $(j \geqslant 1)$, and this contradicts the linear independence of the vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$. Consequently, relation (1) can be solved for $\mathbf{x}$:

$$\mathbf{x} = \gamma_1\mathbf{e}_1 + \gamma_2\mathbf{e}_2 + \ldots + \gamma_n\mathbf{e}_n, \tag{2}$$

where $\gamma_1 = -c_1/c_0$, $\gamma_2 = -c_2c_0$, $\ldots$, $\gamma_n = -c_n/c_0$.

Thus any vector $\mathbf{x}$ of the space U_n is a linear combination of the base vectors.

We shall prove that representation (2) is unique. We assume that there is another representation

$$\mathbf{x} = \gamma_1'\mathbf{e}_1 + \gamma_2'\mathbf{e}_2 + \ldots + \gamma_n'\mathbf{e}_n, \tag{2'}$$

different from representation (2). Subtracting (2′) from (2), we obtain

$$0 = (\gamma_1 - \gamma_1')\,\mathbf{e}_1 + (\gamma_2 - \gamma_2')\,\mathbf{e}_2 - \ldots - (\gamma_n - \gamma_n')\,\mathbf{e}_n. \tag{3}$$

Since the base vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ are linearly independent, there must hold relations

$$(\gamma_1 - \gamma_1') = 0,\ (\gamma_2 - \gamma_2') = 0,\ \ldots,\ (\gamma_n - \gamma_n') = 0,$$

whence it follows that $\gamma_1 = \gamma_1'$, $\gamma_2 = \gamma_2'$, $\ldots$, $\gamma_n = \gamma_n'$, i.e. representation of the vector in terms of the base vectors is unique. ■

The numbers $\gamma_1, \gamma_2, \ldots, \gamma_n$ in relation (2) are the *coordinates of the vector* $\mathbf{x}$ with respect to the basis $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$.

Consider a system of m vectors in the n-dimensional space

$$\begin{aligned}
\mathbf{x}_1 &= a_{11}\mathbf{e}_1 + a_{21}\mathbf{e}_2 + \ldots + a_{n1}\mathbf{e}_n,\\
\mathbf{x}_2 &= a_{12}\mathbf{e}_1 + a_{22}\mathbf{e}_2 + \ldots + a_{n2}\mathbf{e}_n,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
\mathbf{x}_m &= a_{1m}\mathbf{e}_1 + a_{2m}\mathbf{e}_2 + \ldots + a_{nm}\mathbf{e}_n.
\end{aligned}$$

We write the matrix composed of the coordinates of these vectors so that the elements of the jth column are equal to the coordinates of the jth vector ($j = 1, 2, \ldots, m$):

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \cdot & \cdot & \cdot\;\cdot\;\cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nm} \end{pmatrix}.$$

Let the rank of this matrix be equal to r. Then the rth-order minor, different from zero, is a *base minor* of the matrix A. The rows and columns whose intersection is occupied by the base minor are *base rows* and *base columns*.

Theorem 2 (on the base minor). *The base columns (rows) of the matrix A are linearly independent and every column (row) of this matrix is a linear combination of its base columns (rows).*

Example. Given a system of vectors $\mathbf{x}_1 = (-2, 4, 3, 5)$, $\mathbf{x}_2 = (0, 1, 2, -1)$, $\mathbf{x}_3 = (-2, 7, 9, 2)$, we must find the linear dependence of this system of vectors, determine the basis of the system and represent the vectors of the system as a linear combination of the base vectors.

△ (1) We form a matrix from the coordinates of the vectors:

$$A=\begin{bmatrix} -2 & 0 & -2 \\ 4 & 1 & 7 \\ 3 & 2 & 9 \\ 5 & -1 & 2 \end{bmatrix}.$$

Then we determine the rank of the matrix A using the bordering method. We take a minor $M_2 = \begin{vmatrix} -2 & 0 \\ 4 & 1 \end{vmatrix} \neq 0$ and consider the minor bordering it:

$$M_3^1=\begin{vmatrix} -2 & 0 & -2 \\ 4 & 1 & 7 \\ 3 & 2 & 9 \end{vmatrix}.$$

To calculate this minor, we subtract the first column from the third and get

$$M_3^1=\begin{vmatrix} -2 & 0 & 0 \\ 4 & 1 & 3 \\ 3 & 2 & 6 \end{vmatrix}=0$$

since there are two proportional columns.

The second bordering minor

$$M_3^2=\begin{vmatrix} -2 & 0 & -2 \\ 4 & 1 & 7 \\ 5 & -1 & 2 \end{vmatrix}=\begin{vmatrix} -2 & 0 & 0 \\ 4 & 1 & 3 \\ 5 & -1 & -3 \end{vmatrix}=0$$

since, after a similar transformation, it also contains two proportional columns.

Thus all minors of the order higher than the second are zero and therefore the rank of the matrix A is 2 and the minor $M_2 = \begin{vmatrix} -2 & 0 \\ 4 & 1 \end{vmatrix}$ is a base minor. Consequently, the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ are linearly independent and form the basis of the system and the vector $\mathbf{x}_3$ is their linear combination. Thus this system of vectors is linearly dependent.

(2) We set up an equality

$$c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + c_3\mathbf{x}_3 = 0, \qquad (*)$$

whence we get a system of equations

$$\begin{cases} c_1(-2)+c_2\cdot 0+c_3(-2)=0, \\ c_2\cdot 4+c_2\cdot 1+c_3\cdot 7=0. \end{cases}$$

We find the expression for $c_1 = -c_3$ from the first equation and substitute it into the second equation:

$$-4c_3 + c_2 + 7c_3 = 0, \quad c_2 = -3c_3.$$

We set $c_3 = 1$ and then $c_1 = -1$, $c_2 = -3$. Substituting the values of c_1, c_2 and c_3 into relation (*), we get

$$-\mathbf{x}_1 + 3\mathbf{x}_2 + \mathbf{x}_3 = 0, \quad \text{or} \quad \mathbf{x}_3 = \mathbf{x}_1 + 3\mathbf{x}_2.$$

Thus the vector $\mathbf{x}_3$ is represented as a linear combination of the base vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. This representation is unique. The numbers (1, 3) are the coordinates of the vector $\mathbf{x}_3$ in the basis $(\mathbf{x}_1\ \ \mathbf{x}_2)$. ▲

2.14. The Transformation of the Coordinates of a Vector upon a Change in the Basis

Let $\{\mathbf{e}\} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$ and $\{\mathbf{f}\} = (\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n)$ be two bases of the same linear space U_n. Every vector of the new basis $\{\mathbf{f}\}$ has coordinates $s_{1j}, s_{2j}, \ldots, s_{nj}$ in the old basis $\{\mathbf{e}\}$ and in the designation of the coordinates s_{ij} $(i = 1, 2, \ldots, n)$ the first index denotes the number of the old base vector and the second index denotes the number of the corresponding new base vector. Consequently,

$$f_j = s_{1j}\mathbf{e}_1 + s_{2j}\mathbf{e}_2 + \ldots + s_{nj}\mathbf{e}_n \ (j = 1, 2, \ldots, n), \tag{1}$$

or

$$\begin{cases} \mathbf{f}_1 = s_{11}\mathbf{e}_1 + s_{21}\mathbf{e}_2 + \ldots + s_{n1}\mathbf{e}_n, \\ \mathbf{f}_2 = s_{12}\mathbf{e}_1 + s_{22}\mathbf{e}_2 + \ldots + s_{n2}\mathbf{e}_n, \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ \mathbf{f}_n = s_{1n}\mathbf{e}_1 + s_{2n}\mathbf{e}_2 + \ldots + s_{nn}\mathbf{e}_n, \end{cases} \tag{2}$$

where

$$S = \begin{bmatrix} s_{11} & s_{21} & \cdots & s_{n1} \\ s_{12} & s_{22} & \cdots & s_{n2} \\ \cdot & \cdot & \cdot & \cdot \\ s_{1n} & s_{2n} & \cdots & s_{nn} \end{bmatrix}$$

is a nonsingular matrix since $\det S \neq 0$ (otherwise the rows of this matrix and, hence, the vectors $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n$ would be linearly dependent).

The matrix S is the *matrix of the transformation* of the old basis into a new one.

Let **x** be an arbitrary vector of the linear space U_n being considered. We designate the coordinates of this vector in the old basis as x_i and its coordinates in the new basis as y_i. It is evident that

$$\mathbf{x} = y_1\mathbf{f}_1 + y_2\mathbf{f}_2 + \ldots + y_n\mathbf{f}_n = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \ldots + x_n\mathbf{e}_n,$$

or

$$\mathbf{x} = \sum_{i=1}^{n} x_i\mathbf{e}_i = \sum_{j=1}^{n} y_j\mathbf{f}_j.$$

Taking relations (2) into account, we have $\sum\limits_{j=1}^{n} \mathbf{f}_j = \sum\limits_{i=1}^{n} s_{ij}\mathbf{e}_i$, whence, substituting the expression for $\mathbf{f}_j$, we get

$$\mathbf{x} = \sum_{i=1}^{n} x_i\mathbf{e}_i = \sum_{j=1}^{n} y_j \sum_{i=1}^{n} s_{ij}\mathbf{e}_i = \sum_{i=1}^{n} \mathbf{e}_i \sum_{j=1}^{n} s_{ij}y_j. \qquad (3)$$

Consequently, by virtue of the linear independence of the vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$, we find that

$$x_i = \sum_{j=1}^{n} s_{ij}y_j \quad (i = 1, 2, \ldots, n),$$

or

$$\left\{\begin{array}{l} x_1 = s_{11}y_1 + s_{21}y_2 + \ldots + s_{n1}y_n, \\ x_2 = s_{12}y_1 + s_{22}y_2 + \ldots + s_{n2}y_n, \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ x_n = s_{1n}y_1 + s_{2n}y_2 + \ldots + s_{nn}y_n. \end{array}\right. \qquad (4)$$

Relation (4) can be written in matrix form as

$$\mathbf{x} = S\mathbf{y}, \qquad (5)$$

i.e. in the old coordinates (in the old basis) the vector is equal to the transformation matrix S multiplied by the vector in the new coordinates. We premultiply relation (5) by the inverse matrix S^{-1}:

$$S^{-1}\mathbf{x} = \mathbf{y}, \text{ or } \mathbf{y} = S^{-1}\mathbf{x}. \qquad (6$$

The matrix S^{-1} has the form

$$S^{-1} = \begin{bmatrix} S_{11}/d & S_{12}/d & \ldots & S_{1n}/d \\ S_{21}/d & S_{22}/d & \ldots & S_{2n}/d \\ \cdot & \cdot & \cdot & \cdot \\ S_{n1}/d & S_{n2}/d & \ldots & S_{nn}/d \end{bmatrix}$$

where d is the determinant of the matrix S ($d \neq 0$ since S is a nonsingular matrix) and s_{ij} are cofactors of the elements of the determinant d. Then

$$\left\{\begin{array}{l} y_1 = \frac{S_{11}}{d} x_1 + \frac{S_{12}}{d} x_2 + \ldots + \frac{S_{1n}}{d} x_n, \\ y_2 = \frac{S_{21}}{d} x_1 + \frac{S_{22}}{d} x_2 + \ldots + \frac{S_{2n}}{d} x_n, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_n = \frac{S_{n1}}{d} x_1 + \frac{S_{n2}}{d} x_2 + \ldots + \frac{S_{nn}}{d} x_n. \end{array}\right. \tag{7}$$

Relations (7) are formulas for transition from the coordinates x_i of the vector $\mathbf{x}$ in the old basis to the coordinates y_i in the new basis ($i = 1, 2, \ldots, n$). The transition from the coordinates x_i to the coordinates y_i is carried out by means of the matrix S^{-1} which is the inverse of the matrix of the transformation of the old basis into a new one.

Example. Find the coordinates of the vector $\mathbf{x} = (0, 0, 0, 1)$ in the basis $\mathbf{e}_1 = (1, 1, 0, 1)$, $\mathbf{e}_2 = (2, 1, 3, 1)$, $\mathbf{e}_3 = (1, 1, 0, 0)$, $\mathbf{e}_4 = (0, 1, -1, -1)$.

$\triangle$ (1) We compose a matrix S of the transformation of the old basis into the new one:

$$S = \begin{bmatrix} 1 & 2 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 3 & 0 & -1 \\ 1 & 1 & 0 & -1 \end{bmatrix}.$$

(2) We find the inverse matrix

$$S^{-1} = \frac{1}{\det S} \begin{bmatrix} S_{11} & S_{21} & S_{31} & S_{41} \\ S_{12} & S_{22} & S_{32} & S_{42} \\ S_{13} & S_{23} & S_{33} & S_{43} \\ S_{14} & S_{24} & S_{34} & S_{44} \end{bmatrix}.$$

We have

$$\det S = \begin{vmatrix} 1 & 2 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 3 & 0 & -1 \\ 1 & 1 & 0 & -1 \end{vmatrix} = \begin{vmatrix} 1 & 2 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 3 & 0 & -1 \\ 1 & 1 & 0 & -1 \end{vmatrix} = \begin{vmatrix} 0 & -1 & 1 \\ 0 & 3 & -1 \\ 1 & 1 & -1 \end{vmatrix}$$

$$= \begin{vmatrix} -1 & 1 \\ 3 & -1 \end{vmatrix} = -2.$$

We have transformed the determinant of the matrix S in the following way: first we subtracted the first row from the second, then expanded the determinant obtained according to the elements of the third column, and finally expanded the resulting determinant according to the elements of the first column.

We calculate the cofactors:

$$S_{11}=\begin{vmatrix}1&1&1\\3&0&-1\\1&0&-1\end{vmatrix}=2;\quad S_{12}=-\begin{vmatrix}1&1&1\\0&0&-1\\1&0&-1\end{vmatrix}=1,$$

$$S_{13}=\begin{vmatrix}1&1&1\\0&3&-1\\1&1&-1\end{vmatrix}=\begin{vmatrix}1&1&1\\0&3&-1\\0&0&-2\end{vmatrix}=-6,$$

$$S_{14}=-\begin{vmatrix}1&1&1\\0&3&0\\1&1&0\end{vmatrix}=-\begin{vmatrix}0&3\\1&1\end{vmatrix}=3,$$

$$S_{21}=-\begin{vmatrix}2&1&0\\3&0&-1\\1&0&-1\end{vmatrix}=\begin{vmatrix}3&-1\\1&-1\end{vmatrix}=-2,$$

$$S_{22}=\begin{vmatrix}1&1&0\\0&0&-1\\1&0&-1\end{vmatrix}=-\begin{vmatrix}0&-1\\1&-1\end{vmatrix}=-1,$$

$$S_{23}=-\begin{vmatrix}1&2&0\\0&3&-1\\1&1&-1\end{vmatrix}=-\begin{vmatrix}1&2&0\\0&3&-1\\0&-1&-1\end{vmatrix}=-\begin{vmatrix}3&-1\\-1&-1\end{vmatrix}=-4,$$

$$S_{24}=\begin{vmatrix}1&2&1\\0&3&0\\1&1&0\end{vmatrix}=\begin{vmatrix}0&3\\1&1\end{vmatrix}=-3,$$

$$S_{31}=\begin{vmatrix}2&1&0\\1&1&1\\1&0&-1\end{vmatrix}=0$$ (the first row is equal to the sum of the second and the third row),

$$S_{32}=-\begin{vmatrix}1&1&0\\1&1&1\\1&0&-1\end{vmatrix}=-\begin{vmatrix}1&1&0\\0&0&1\\1&0&-1\end{vmatrix}=\begin{vmatrix}0&1\\1&-1\end{vmatrix}=-1,$$

$$S_{33}=\begin{vmatrix}1&2&0\\1&1&1\\1&1&-1\end{vmatrix}=\begin{vmatrix}1&2&0\\1&1&1\\2&2&0\end{vmatrix}=-\begin{vmatrix}1&2\\2&2\end{vmatrix}=2,$$

$$S_{34}=-\begin{vmatrix}1&2&1\\1&1&1\\1&1&0\end{vmatrix}=-\begin{vmatrix}1&2&1\\0&-1&0\\1&1&0\end{vmatrix}=-\begin{vmatrix}0&-1\\1&1\end{vmatrix}=-1,$$

$$S_{41}=-\begin{vmatrix}2&1&0\\1&1&1\\3&0&-1\end{vmatrix}=-\begin{vmatrix}2&1&0\\1&1&1\\4&1&0\end{vmatrix}=\begin{vmatrix}2&1\\4&1\end{vmatrix}=-2,$$

$$S_{42}=\begin{vmatrix}1 & 1 & 0\\ 1 & 1 & 1\\ 0 & 0 & -1\end{vmatrix}=0 \text{ (there are two identical columns),}$$

$$S_{43}=-\begin{vmatrix}1 & 2 & 0\\ 1 & 1 & 1\\ 0 & 3 & -1\end{vmatrix}=-\begin{vmatrix}1 & 2 & 0\\ 1 & 4 & 0\\ 0 & 3 & -1\end{vmatrix}=\begin{vmatrix}1 & 2\\ 1 & 4\end{vmatrix}=2,$$

$$S_{44}=\begin{vmatrix}1 & 2 & 1\\ 1 & 1 & 1\\ 0 & 3 & 0\end{vmatrix}=0.$$

Consequently,

$$S^{-1}=\frac{1}{2}\begin{vmatrix}2 & -2 & 0 & -2\\ 1 & -1 & -1 & 0\\ -6 & 4 & 2 & 2\\ 3 & -3 & -1 & 0\end{vmatrix}.$$

(3) From formula (6) we find the coordinates of the vector $\mathbf{x}$ in the new basis:

$$\mathbf{y}=S^{-1}\mathbf{x}=\frac{1}{2}\begin{bmatrix}2 & -2 & 0 & -2\\ 1 & -1 & -1 & 0\\ -6 & 4 & 2 & 2\\ 3 & -3 & -1 & 0\end{bmatrix}\begin{bmatrix}0\\ 0\\ 0\\ 1\end{bmatrix}$$

$$=\frac{1}{2}\begin{bmatrix}-2\\ 0\\ 2\\ 0\end{bmatrix}=\begin{bmatrix}1\\ 0\\ -1\\ 0\end{bmatrix}. \ ▲$$

Exercises

1. Calculate AB if

(a) $A=\begin{bmatrix}1 & -2 & 3\\ 3 & -1 & 2\\ 4 & -2 & 1\end{bmatrix}$, $B=\begin{bmatrix}2 & 3 & 1\\ 1 & 2 & 3\\ 2 & 1 & 3\end{bmatrix}$,

(b) $A=\begin{bmatrix}1 & -3 & 2\\ 3 & -4 & 1\\ 2 & -5 & 3\end{bmatrix}$, $B=\begin{bmatrix}2 & 5 & 6\\ 1 & 2 & 5\\ 1 & 3 & 2\end{bmatrix}$.

2. Calculate $2(A+B)(2B-A)$ if

(a) $A=\begin{bmatrix}2 & 3 & -1\\ 4 & 5 & 2\\ -1 & 0 & 7\end{bmatrix}$, $B=\begin{bmatrix}-1 & 0 & 5\\ 0 & 1 & 3\\ 2 & -2 & 4\end{bmatrix}$,

(b) $A=\begin{bmatrix}4 & 5 & -2\\ 3 & -1 & 0\\ 4 & 2 & 7\end{bmatrix}$, $B=\begin{bmatrix}2 & 1 & -1\\ 0 & 1 & 3\\ 5 & 7 & 3\end{bmatrix}$.

3. Find the product XY if

(a) $X=\begin{bmatrix}5\\7\\-3\\2\end{bmatrix}$, $Y=[1\ 2\ -2\ -3]$, (b) $X=\begin{bmatrix}1\\2\\3\end{bmatrix}$, $Y=[4\ 5]$,

(c) $X=[10\ 17\ 8\ 5\ 11]$, $Y=\begin{bmatrix}12\\7\\5\\4\\10\end{bmatrix}$.

4. Find the product AX if

(a) $A=\begin{bmatrix}2 & -1 & -3 & 0 & 4\\7 & 2 & 5 & 5 & 2\\3 & 4 & 1 & -7 & -1\end{bmatrix}$, $X=\begin{bmatrix}-1\\1\\2\\3\\0\end{bmatrix}$.

(b) $A=\begin{bmatrix}1 & 2 & -4\\3 & 0 & -2\\2 & 2 & -3\end{bmatrix}$, $X=\begin{bmatrix}-2\\-3\\-4\end{bmatrix}$.

5. Calculate the determinants

(a) $d=\begin{bmatrix}1 & 3 & -4\\0 & 1 & 1\\2 & -5 & 3\end{bmatrix}$, (b) $d=\begin{vmatrix}1 & 1 & -2 & 3\\7 & 8 & 4 & 1\\2 & 4 & 6 & -3\\5 & 6 & 8 & -4\end{vmatrix}$,

(c) $d=\begin{vmatrix}1.6 & 5.4 & -7.7 & -3.1\\8.2 & 1.4 & -2.3 & 0.2\\5.3 & -5.9 & 2.7 & -7.9\\0.7 & 1.9 & -8.5 & 4.8\end{vmatrix}$.

6. Calculate A^{-1} for the following matrices:

(a) $A=\begin{bmatrix}1 & -3 & 2\\3 & -4 & 0\\2 & -5 & 3\end{bmatrix}$, (b) $A=\begin{bmatrix}2 & 7 & 1 & 4\\5 & 2 & 0 & -1\\3 & 4 & 2 & 1\\6 & 8 & 4 & 3\end{bmatrix}$,

(c) $A=\begin{bmatrix}1 & 0 & 0 & 0\\0 & 2 & 0 & 0\\2 & -3 & 4 & 0\\3 & 2 & -1 & 3\end{bmatrix}$.

7. Find AB, where

$$A=\begin{bmatrix} 1 & 4 & 1 & 3 \\ 0 & -1 & 3 & -1 \\ 3 & 1 & 0 & 2 \\ 1 & -2 & 5 & 1 \end{bmatrix},\quad B=\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 4 & 2 & 3 \\ 1 & 10 & 3 & 6 \\ 6 & 10 & 1 & 4 \end{bmatrix}.$$

using two techniques: (a) by partitioning A and B into square blocks, (b) by partitioning A and B into blocks by bordering.

8. Calculate A^{-1} using the partitioning into blocks and bordering if

$$\text{(a) } A=\begin{bmatrix} 1 & 2 & 3 & -2 \\ 2 & -1 & -2 & -3 \\ 3 & 2 & -1 & 2 \\ 2 & -3 & 2 & 1 \end{bmatrix},\quad \text{(b) } A=\begin{bmatrix} 3 & -2 & 2 & 0 \\ 2 & 1 & 1 & -2 \\ 3 & -1 & 2 & 1 \\ 1 & 2 & -1 & -1 \end{bmatrix}.$$

9. Expand the matrices given in Exercise 8 in the product of two triangular matrices and invert them using the expansion of matrices in the product of two triangular matrices.

10. Solve the following matrix equations:

$$\text{(a) } X\begin{bmatrix} 0 & 3 & -1 \\ 2 & -1 & 2 \\ -3 & 1 & 4 \end{bmatrix}=\begin{bmatrix} 7 & 6 & -3 \\ -8 & 3 & 6 \\ 11 & 9 & 13 \end{bmatrix},$$

$$\text{(b) } \begin{bmatrix} 7 & 6 & -3 \\ -8 & 3 & 6 \\ 11 & 9 & 13 \end{bmatrix}X=\begin{bmatrix} -3 & -10 & -4 \\ 21 & 14 & -10 \\ 48 & 2 & 30 \end{bmatrix}.$$

11. Calculate the ranks of the following matrices:

$$\text{(a) } A=\begin{bmatrix} 2 & -1 & 3 & -2 & 4 \\ 4 & -2 & 5 & 1 & 7 \\ 2 & -1 & 1 & 8 & 2 \end{bmatrix},\quad \text{(b) } B=\begin{bmatrix} 2 & 4 & 1 & 0 \\ 2 & 4 & 1 & 0 \\ -1 & -2 & 3 & 1 \\ 5 & 10 & 6 & 1 \\ 0 & 0 & 7 & 2 \end{bmatrix}.$$

12. Test the following systems of vectors for linear dependence: (a) $\mathbf{x}_1 = (5, 4, 3)$, $\mathbf{x}_2 = (3, 3, 2)$, $\mathbf{x}_3 = (8, 1, 3)$, (b) $\mathbf{x}_1 = (1, 0, 0, 2, 5)$, $\mathbf{x}_2 = (0, 1, 0, 3, 4)$, $\mathbf{x}_3 = (0, 0, 1, 4, 7)$, $\mathbf{x}_4 = (2, -3, 4, 11, 12)$.

13. For the system of vectors $\mathbf{x}_1 = (5, 2, -3, 1)$, $\mathbf{x}_2 = (4, 1, -2, 3)$, $\mathbf{x}_3 = (1, 1, -1, -2)$, $\mathbf{x}_4 = (3, 4, -1, 2)$ find the basis and express the other vectors in terms of the base vectors.

14. Find the coordinates of the vector $\mathbf{x} = (1, 2, 1, 1)$ in the basis $\mathbf{e}_1 = (1, 1, 1, 1)$, $\mathbf{e}_2 = (1, 1, -1, -1)$, $\mathbf{e}_3 = (1, -1, 1, -1)$, $\mathbf{e}_4 = (1, -1, -1, 1)$.

Chapter 3

Solving Systems of Linear Equations

3.1. Systems of Linear Equations

In the general form a *system of m linear equations in n unknowns* is written as

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \dots + a_{1j}x_j + \dots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2j}x_j + \dots + a_{2n}x_n = b_2, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ a_{i1}x_1 + a_{i2}x_2 + \dots + a_{ij}x_j + \dots + a_{in}x_n = b_i, \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mj}x_j + \dots + a_{mn}x_n = b_m. \end{cases} \qquad (1)$$

The equations of the system are assumed to be enumerated, i.e. the first, the second, . . ., the *m*th. The numbers $x_1, x_2, \dots, x_n$ are the *unknowns* of the system and $a_{11}, a_{12}, \dots, a_{mn}$ are the *coefficients* of the unknowns of the system.

The coefficient of the unknown x_{ij} in the *i*th equation is designated as a_{ij}, where the first index i indicates the number of the equation which contains this coefficient and the second index j indicates the number of the unknown in which this coefficient is. For instance, the coefficient a_{23} is in the second equation of the system in the unknown x_3.

The numbers $b_1, b_2, \dots, b_m$ are *constant*, or *free, terms* of the system.

In the abbreviated form system (1) can be written as

$$\sum_{j=i}^{n} a_{ij}x_j = b_i \quad (i = 1, 2, \dots, m). \qquad (1')$$

A *solution* of the system of linear equations (1) is any set of numbers $\alpha_1, \alpha_2, \dots, \alpha_n$, which, being substituted for the unknowns $x_1, x_2, \dots, x_n$ into the equations of the system, turn all the equations into identities.

The system of linear equations (1) is *consistent* if it has a solution. If a system of linear equations has no solution, then it is *inconsistent* (or *incompatible*).

A consistent system of linear equations may have one solution or several solutions and is said to be *determinate* if there is one solution and *indeterminate* if there are more than one solution.

Two systems of linear equations with the same number of unknowns are *equivalent* if they are either both inconsistent or both consistent and have the same solutions.

Here are three types of the *elementary transformations of a system of linear equations*:

(1) permutation of two equations of the system,

(2) multiplication of both sides of an equation of the system by any nonzero number,

(3) addition to (subtraction from) both sides of one equation of the corresponding sides of another equation multiplied by any number.

We can prove that elementary transformations turn a given system of equations into an equivalent system. To perform elementary transformations is the same as to express one unknown in terms of the others.

A system in which the constant terms b_1, b_2, . . ., b_n are zero is a *homogeneous* system.

3.2. The Kronecker-Capelli Theorem

Consider a system of linear equations

$$\left\{\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2, \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n = b_m. \end{array}\right. \quad (1)$$

To establish the conditions of consistency of the system, it is necessary to introduce the concept of the matrix of a system and the augmented matrix of a system.

The *matrix of system* (1) is a matrix composed of the coefficients of the unknowns of the system:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdot & a_{1n} \\ a_{21} & a_{22} & \cdot\cdot & a_{2n} \\ \cdot & \cdot & \cdot\cdot & \cdot \\ a_{m1} & a_{m2} & & a_{mn} \end{bmatrix}.$$

If we add a column of constant terms to the matrix A, we get a matrix $\bar{A}$ which is known as the *augmented matrix of system* (1):

$$\bar{A}=\begin{bmatrix} a_{11} & a_{12} & \dots a_{1n} b_1 \\ a_{21} & a_{22} & \dots a_{2n} b_2 \\ \cdot & \cdot & \cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot \\ a_{m1} & a_{m2} & \dots a_{mn} b_m \end{bmatrix}.$$

It is clear from the definition of the system matrix A and the augmented matrix $\bar{A}$ that their ranks $r(A)$ and $r(\bar{A})$ are either equal or the rank $r(\bar{A})$ is larger by unity than $r(A)$.

The question concerning the consistency of system (1) is answered by the Kronecker-Capelli theorem: *the system of linear equations* (1) *is consistent if and only if the rank of the augmented matrix $\bar{A}$ is equal to the rank of the matrix A, i.e. when $r(\bar{A}) = r(A)$.*

Example. Test the following system of linear equations for consistency:

$$\begin{cases} 7x_1+3x_2=2, \\ x_1-2x_2=-3, \\ 4x_1+9x_2=11. \end{cases}$$

△ (1) We set up a matrix for the given system and calculate its rank:

$$A=\begin{bmatrix} 7 & 3 \\ 1 & -2 \\ 4 & 9 \end{bmatrix}, \quad r(A)=2, \quad \text{since } M_2^1=\begin{vmatrix} 7 & 3 \\ 1 & -2 \end{vmatrix} \neq 0.$$

(2) We set up an augmented matrix for the system:

$$\bar{A}=\begin{bmatrix} 7 & 3 & 2 \\ 1 & -2 & -3 \\ 4 & 9 & 11 \end{bmatrix}.$$

Since $M_2^1 \neq 0$ and the minor bordering it

$$M_3=\det \bar{A}=\begin{vmatrix} 7 & 3 & 2 \\ 1 & -2 & -3 \\ 4 & 9 & 11 \end{vmatrix}=0$$

(the first row is equal to the sum of the second row multiplied by 3 and the third row), it follows that $r(\bar{A}) = 2$. Thus $r(A) = r(\bar{A}) = 2$, i.e. the system is consistent. ▲

Corollary 1. *If system* (1) *is consistent and the rank of the system matrix* $r(A) = r$ *is equal to the number of unknowns* n, *then the system has a unique solution.*

Corollary 2. *If system* (1) *is consistent and the rank of the system matrix* $r(A) = r$ *is lower than the number of the unknowns* n, *then the system has an infinite number of solutions.*

3.3. Cramer's Rule for *n* Linear Equations in *n* Unknowns

Consider a system of linear equations in which the number of equations is equal to the number of unknowns

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n, \end{cases} \tag{1}$$

and

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

are the system matrix, the column of constant terms and the column of unknowns respectively. We assume that the determinant of the system

$$d = \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{vmatrix} \neq 0.$$

If now we successively replace in the determinant d the columns of the coefficients of the unknowns x_j ($j = 1, 2, \ldots, n$) by the column of constant terms b_i, we obtain determinants

$$d_1 = \begin{vmatrix} b_1 & a_{12} & \ldots & a_{1,n-1} & a_{1n} \\ b_2 & a_{22} & \ldots & a_{2,n-1} & a_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ b_n & a_{n2} & \ldots & a_{n,n-1} & a_{nn} \end{vmatrix},$$

$$d_2 = \begin{vmatrix} a_{11} & b_1 & \dots & a_{1,n-1} & a_{1n} \\ a_{21} & b_2 & \dots & a_{2,n-1} & a_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & b_n & \dots & a_{n,n-1} & a_{nn} \end{vmatrix},$$

$$\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot$$

$$d_{n-1} = \begin{vmatrix} a_{11} & a_{12} & \dots & b_1 & a_{1n} \\ a_{21} & a_{22} & \dots & b_2 & a_{2n} \\ a_{n1} & a_{n2} & \dots & b_n & a_{nn} \end{vmatrix},$$

$$d_n = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1,n-1} & b_1 \\ a_{21} & a_{22} & \dots & a_{2,n-1} & b_2 \\ a_{n1} & a_{n2} & \dots & a_{n,n-1} & b_n \end{vmatrix},$$

respectively.

Cramer's rule. *A system of n linear equations in n unknowns, whose determinant is different from zero, is always consistent and has a unique solution which can be found from the formulas*

$$x_1 = d_1/d,\ x_2 = d_2/d,\ \dots,\ x_{n-1} = d_{n-1}/d,$$
$$x_n = d_n/d. \qquad (2)$$

Formulas (2) are known as *Cramer's formulas.*

Example 1. Use Cramer's formulas to solve the following system of linear equations:

$$\begin{cases} x_1 + x_2 + 2x_3 = -1, \\ 2x_1 - x_2 + 2x_3 = -4, \\ 4x_1 + x_2 + 4x_3 = -2. \end{cases}$$

△ (1) We calculate the determinant of the system:

$$d = \begin{vmatrix} 1 & 1 & 2 \\ 2 & -1 & 2 \\ 4 & 1 & 4 \end{vmatrix} = 2 \begin{vmatrix} 1 & 1 & 1 \\ 2 & -1 & 1 \\ 4 & 1 & 2 \end{vmatrix} = 2(-2+4+2+4-4-1) = 6.$$

(2) We calculate the determinants composed of the coefficients of the unknowns x_1, x_2, x_3:

$$d_1 = \begin{vmatrix} -1 & 1 & 2 \\ -4 & -1 & 2 \\ -2 & 1 & 4 \end{vmatrix} = 2 \begin{vmatrix} -1 & 1 & 1 \\ -4 & -1 & 1 \\ -2 & 1 & 2 \end{vmatrix}$$
$$= 2(2-2-4-2+1+8) = 6,$$

$$d_2=\begin{vmatrix}1 & -1 & 2\\ 2 & -4 & 2\\ 4 & -2 & 4\end{vmatrix}=2\begin{vmatrix}1 & -1 & 1\\ 2 & -4 & 1\\ 4 & -2 & 2\end{vmatrix}=4\begin{vmatrix}1 & -1 & 1\\ 2 & -4 & 1\\ 2 & -1 & 1\end{vmatrix}$$

$$=4(-4-2-2+8+2+1)=12,$$

$$d_3=\begin{vmatrix}1 & 1 & -1\\ 2 & -1 & -4\\ 4 & 1 & -2\end{vmatrix}=(2-16-2-4+4+4)=-12.$$

(3) Using Cramer's formulas (2), we find the solution of the system:

$$x_1=d_1/d=6/6=1,\quad x_2=d_2/d=12/6=2,\quad x_3=-12/6=-2.\ \blacktriangle$$

Example 2. Use Cramer's formulas to solve the following system of linear equations:

$$\begin{cases}2x_1-x_2+x_3+3x_4=-1,\\ x_1+x_2-x_3-4x_4=6,\\ 3x_1-x_2+x_3+x_4=4,\\ x_1-3x_2+3x_4=-5.\end{cases}$$

$\triangle$ We find the determinants d, d_1, d_2, d_3 and d_4 expanding them into minors according to the elements of the last row and then applying the rule of triangles:

$$d=\begin{vmatrix}2 & -1 & 1 & 3\\ 1 & 1 & -1 & -4\\ 3 & -1 & 1 & 1\\ 1 & -3 & 0 & 3\end{vmatrix}=-\begin{vmatrix}-1 & 1 & 3\\ 1 & -1 & -4\\ -1 & 1 & 1\end{vmatrix}$$

$$-3\begin{vmatrix}2 & 1 & 3\\ 1 & -1 & -4\\ 3 & 1 & 1\end{vmatrix}+3\begin{vmatrix}2 & -1 & 1\\ 1 & 1 & -1\\ 3 & -1 & 1\end{vmatrix}=(-1)\cdot 0-3\cdot 5+3\cdot 0=-15$$

(the first and the third determinant are zero since they have proportional columns),

$$d_1=\begin{vmatrix}-1 & -1 & 1 & 3\\ 6 & 1 & -1 & -4\\ 4 & -1 & 1 & 1\\ -5 & -3 & 0 & 3\end{vmatrix}=5\begin{vmatrix}-1 & 1 & 3\\ 1 & -1 & -4\\ -1 & 1 & 1\end{vmatrix}$$

$$-3\begin{vmatrix}-1 & 1 & 3\\ 6 & -1 & -4\\ 4 & 1 & 1\end{vmatrix}+3\begin{vmatrix}-1 & -1 & 1\\ 6 & 1 & -1\\ 4 & -1 & 1\end{vmatrix}=5\cdot 0-3\cdot 5+3\cdot 0=-15.$$

$$d_2=\begin{vmatrix}2 & -1 & 1 & 3\\ 1 & 6 & -1 & -4\\ 3 & 4 & 1 & 1\\ 1 & -5 & 0 & 3\end{vmatrix}=-\begin{vmatrix}-1 & 1 & 3\\ 6 & -1 & -4\\ 4 & 1 & 1\end{vmatrix}$$

$$-5\begin{vmatrix}2&1&3\\1&-1&-4\\3&1&1\end{vmatrix}+3\begin{vmatrix}2&-1&1\\1&6&-1\\3&4&1\end{vmatrix}=-1\cdot5-5\cdot5+3\cdot10=0,$$

$$d_3=\begin{vmatrix}2&-1&-1&3\\1&1&6&-4\\3&-1&4&1\\1&-3&-5&3\end{vmatrix}=-\begin{vmatrix}-1&-1&3\\1&6&-4\\-1&4&1\end{vmatrix}$$

$$-3\begin{vmatrix}2&-1&3\\1&6&-4\\3&4&1\end{vmatrix}+5\begin{vmatrix}2&-1&3\\1&1&-4\\3&-1&1\end{vmatrix}+3\begin{vmatrix}2&-1&1\\1&1&6\\3&-1&4\end{vmatrix}$$

$$=(-1)\cdot5-3\cdot15+5\cdot(-5)+3\cdot10=-45.$$

$$d_4=\begin{vmatrix}2&-1&1&-1\\1&1&-1&6\\3&-1&1&4\\1&-3&0&-5\end{vmatrix}=-\begin{vmatrix}-1&1&-1\\1&-1&6\\-1&1&4\end{vmatrix}$$

$$-3\begin{vmatrix}2&1&-1\\1&-1&6\\3&1&4\end{vmatrix}-5\begin{vmatrix}2&-1&1\\1&1&-1\\3&-1&1\end{vmatrix}$$

$$=(-1)\cdot0-3\cdot(-10)-5\cdot0=30.$$

Using now Cramer's formulas (2), we get the solution of the system:

$$x_1=d_1/d=(-15)/(-15)=1,\quad x_2=d_2/d=0/(-15)=0,$$
$$x_3=d_3/d=(-45)/(-15)=3,\quad x_4=d_4/d=30/(-15)=-2.\ \blacktriangle$$

Note that the solution of a system of linear equations via Cramer's formulas is very cumbersome. In practical calculations other methods are usually used to solve systems of this kind.

3.4. Solving Arbitrary Systems of Linear Equations

Let

$$\begin{cases}a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n=b_1,\\a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n=b_2,\\\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\a_{m1}x_1+a_{m2}x_2+\ldots+a_{mn}x_n=b_m\end{cases}\qquad(1)$$

be an arbitrary system of linear equations, where the number m of equations is not equal to the number n of unknowns ($m\neq n$).

We assume that system (1) is consistent, i.e. $r(A) = r(\bar{A}) = r$ and $r \leqslant \min\{m, n\}$. Then in the matrices A and $\bar{A}$ of the system there are r linearly independent rows and the other $m - r$ rows are their linear combinations. Interchanging the equations, we may attain a situation when these r linearly independent rows occupy the first r places.

It follows that any one of the last $m - r$ equations of system (1) can be represented as the sum of the first r equations (which are *linearly independent* or *base equations*), taken with certain coefficients. Then system (1) is equivalent to the following system of r equations in n unknowns:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = b_2, \\ \dots\dots\dots\dots\dots\dots\dots \\ a_{r1}x_1 + a_{r2}x_2 + \dots + a_{rn}x_n = b_r \end{cases} \tag{2}$$

We assume that the rth-order minor composed of the coefficients of the first r unknowns is nonzero:

$$M_r = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1r} \\ a_{21} & a_{22} & \dots & a_{2r} \\ \cdot & \cdot & \cdot & \cdot \\ a_{r1} & a_{r2} & \dots & a_{rr} \end{vmatrix} \neq 0,$$

i.e. is a base minor (see 2.13). In that case the unknowns whose coefficients constitute a base minor are *base unknowns* and the other $n - r$ unknowns are *constant unknowns*.

In each of the equations of system (2) we transfer all terms with constant unknowns $x_{r+1}, x_{r+2}, \dots, x_n$ to the right-hand side. Then we get a system

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \dots + a_{1r}x_r \\ \quad = b_1 - a_{1,r+1}x_{r+1} - a_{1,r+2}x_{r+2} - \dots - a_{1n}x_n, \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2r}x_r \\ \quad = b_2 - a_{2,r+1}x_{r+1} - a_{2,r+2}x_{r+2} - \dots - a_{2n}x_n, \\ \dots\dots\dots\dots\dots\dots\dots \\ a_{r1}x_1 + a_{r2}x_2 + \dots + a_{rr}x_r \\ \quad = b_r - a_{r,r+1}x_{r+1} - a_{r,r+1}x_{r+2} - \dots - a_{rn}x_n, \end{cases} \tag{3}$$

which contains r equations in r base unknowns. Since the determinant of system (3) is a base minor $M_r \neq 0$, system (3) has a unique solution for the base unknowns $x_1, x_2, \ldots, x_r$ which can be found from Cramer's formulas. Assigning arbitrary numerical values $x_{r+1} = c_1, x_{r+2} = c_2, \ldots, x_n = c_{n-r}$ to the constant unknowns $x_{r+1}, x_{r+2}, \ldots, x_n$, we get a special solution of the original system (1).

Example. Solve the system of three equations in four unknowns

$$\begin{cases} 3x_1 - 2x_2 + 5x_3 + 4x_4 = 2, \\ 6x_1 - 4x_2 + 4x_3 + 3x_4 = 3, \\ 9x_1 - 6x_2 + 3x_3 + 2x_4 = 4. \end{cases} \qquad (*)$$

△ We test the system for consistency, for which purpose we compose matrices A and $\bar{A}$:

$$A = \begin{bmatrix} 3 & -2 & 5 & 4 \\ 6 & -4 & 4 & 3 \\ 9 & -6 & 3 & 2 \end{bmatrix}, \quad \bar{A} = \begin{bmatrix} 3 & -2 & 5 & 4 & 2 \\ 6 & -4 & 4 & 3 & 3 \\ 9 & -6 & 3 & 2 & 4 \end{bmatrix}.$$

Then we determine the ranks of these matrices using elementary transformations:

(a) since the first and the second column are proportional, we delete one of them (the second),

(b) we multiply the first column by 1/3,

(c) we multiply the result in succession, by (-5) and (—4) and add to the second and the third column respectively,

(d) in the matrix obtained, two columns (the second and the third) are proportional; we delete one of them (the third) and multiply the second column by —1/6,

(e) we multiply, in succession, the first row by (—2) and (—3) and add the result to the second and the third row respectively,

(f) we multiply the second row by (—2), add the result to the third row and delete the zero row obtained.

We have

$$A = \begin{bmatrix} 3 & -2 & 5 & 4 \\ 6 & -4 & 4 & 3 \\ 9 & -6 & 3 & 2 \end{bmatrix} \overset{(a)}{\sim} \begin{bmatrix} 3 & 5 & 4 \\ 6 & 4 & 3 \\ 9 & 3 & 2 \end{bmatrix} \overset{(b)}{\sim} \begin{bmatrix} 1 & 5 & 4 \\ 2 & 4 & 3 \\ 3 & 3 & 2 \end{bmatrix}$$

$$\overset{(c)}{\sim} \begin{bmatrix} 1 & 0 & 0 \\ 2 & -6 & -5 \\ 3 & -12 & -10 \end{bmatrix} \overset{(d)}{\sim} \begin{bmatrix} 1 & 0 \\ 2 & 1 \\ 3 & 2 \end{bmatrix} \overset{(e)}{\sim} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 2 \end{bmatrix} \overset{(f)}{\sim} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Evidently, the rank of the last matrix is 2: $r(A) = 2$.

We similarly transform the matrix $\bar{A}$:

$$\bar{A} = \begin{bmatrix} 3 & -2 & 5 & 4 & 2 \\ 6 & -4 & 4 & 3 & 3 \\ 9 & -6 & 3 & 2 & 4 \end{bmatrix} \sim \begin{bmatrix} 3 & 5 & 4 & 2 \\ 6 & 4 & 3 & 3 \\ 9 & 3 & 2 & 4 \end{bmatrix} \sim \begin{bmatrix} 1 & 5 & 4 & 2 \\ 2 & 4 & 3 & 3 \\ 3 & 3 & 2 & 4 \end{bmatrix}$$

$$\sim \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & -6 & -5 & -1 \\ 3 & -12 & -10 & -2 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 \\ 2 & -1 \\ 3 & -2 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 \\ 0 & -1 \\ 0 & -2 \end{bmatrix}$$

$$\sim \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Consequently, $r(\bar{A}) = 2$.

Thus $r(A) = r(\bar{A}) = 2$, i.e. the system is consistent.

Since the rank of the system is 2, it follows that the maximum order of the nonzero minor is 2 and the system has two base unknowns. We find some nonzero second-order minor, such as, for instance, a minor $M_2 = \begin{vmatrix} 5 & 4 \\ 4 & 3 \end{vmatrix} \neq 0$, formed by the coefficients of the unknowns x_3 and x_4. Consequently, x_3 and x_4 can be considered to be base unknowns and x_1 and x_2 to be constant unknowns.

System (*) is equivalent to the following system:

$$\left\{ \begin{array}{l} 3x_1 - 2x_2 + 5x_3 + 4x_4 = 2, \\ 6x_1 - 4x_2 + 4x_3 + 3x_4 = 3. \end{array} \right. \qquad (**)$$

We transfer the constant unknowns to the right-hand side

$$\left\{ \begin{array}{l} 5x_3 + 4x_4 = 2 - 3x_1 + 2x_2, \\ 4x_3 + 3x_4 = 3 - 6x_1 + 4x_2. \end{array} \right. \qquad (***)$$

and solve system (3) using Cramer's formulas:

$$d = \begin{vmatrix} 5 & 4 \\ 4 & 3 \end{vmatrix} = 15 - 16 = -1,$$

$$d_3 = \begin{vmatrix} 2 - 3x_1 + 2x_2 & 4 \\ 3 - 6x_1 + 4x_2 & 3 \end{vmatrix} = 3(2 - 3x_1 + 2x_2) - 4(3 - 6x_1 + 4x_2)$$

$$= -6 + 15x_1 - 10x_2,$$

$$d_4 = \begin{vmatrix} 5 & 2 - 3x_1 + 2x_2 \\ 4 & 3 - 6x_1 + 4x_2 \end{vmatrix} = 5(3 - 6x_1 + 4x_2) - 4(2 - 3x_1 + 2x_2)$$

$$= 7 - 18x_1 + 12x_2,$$

$$x_3 = d_3/d = 6 - 15x_1 + 10x_2, \quad x_4 = d_4/d = -7 + 18x_1 - 12x_2.$$

The solution obtained, in which the base unknowns x_3 and x_4 are expressed in terms of the constant unknowns x_1 and x_2, is the general solution of system (*). Substituting the arbitrary values of the constant unknowns into it, we get various special solutions. For instance, if $x_1 = 0$, $x_2 = 0$, then $x_3 = 6$, $x_4 = -7$, if $x_1 = 1$, $x_2 = 2$, then $x_3 = 11$, $x_4 = 13$, etc. The collections of numbers (0, 0, 6, −7), (1, 2, 11, 13) etc. are special solutions of system (*). ▲

3.5. Homogeneous Systems of Linear Equations

Consider a homogeneous system of m linear equations in n unknowns:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = 0, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n = 0. \end{cases} \qquad (1)$$

Since the addition of a column of zeros does not change the rank of the matrix of the system, this system is always consistent by virtue of the Kronecker-Capelly theorem and has at least a zero solution ($x_1 = x_2 = \ldots = x_n = 0$). If the determinant of system (1) is nonzero and the number of equations of the system is equal to the number of the unknowns, then, according to Cramer's rule, the zero solution is unique.

In the case when the rank of the matrix of system (1) is lower than the number of the unknowns, i.e. $r(A) < n$, the system has nonzero solutions in addition to the zero one. To find these solutions, we isolate in system (1) r linearly independent equations and delete the other equations. On the left-hand side of the isolated equations we leave r base unknowns and transfer the other $n - r$ constant unknowns to the right-hand side. Then we arrive at a system

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1r}x_r \\ = -a_{1,r+1}x_{r+1} - a_{1,r+2}x_{r+2} - \ldots - a_{1n}x_n, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2r}x_r \\ = -a_{2,r+1}x_{r+1} - a_{2,r+2}x_{r+2} - \ldots - a_{2n}x_n, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{r1}x_1 + a_{r2}x_2 + \ldots + a_{2r}x_r \\ = -a_{r,r+1}x_{r+1} - a_{r,r+2}x_{r+2} - \ldots - a_{rn}x_n, \end{cases} \qquad (2)$$

solving which by Cramer's formulas, we express r base unknowns $x_1, x_2, \ldots, x_r$ in terms of $n - r$ constant unknowns $x_{r+1}, x_{r+2}, \ldots, x_n$.

System (1) has infinitely many solutions, among which there are linearly independent solutions.

A *fundamental system of solutions* consists of $n - r$ linearly independent solutions of a homogeneous system of equations.

Example. Given a homogeneous system of equations

$$\begin{cases} 2x_1-4x_2+\ \ 5x_3+3x_4\ =0, \\ 3x_1-6x_2+\ \ 4x_3+2x_4\ =0, \\ 4x_1-8x_2+17x_3+11x_4=0, \end{cases}$$

find its general solution and the fundamental system of solutions.

△ (1) The number of unknowns here $n = 4$, the number of equations $m = 3$. We calculate the rank of the system matrix using elementary transformations:

(a) delete the second column since it is proportional to the first column,

(b) multiply the third column by (—2) and add the result to the second column and then multiply it by (—3) and add it to the first column multiplied by 2,

(c) delete the first column since it is proportional to the second,

(d) multiply the first column by 3 and add the result to the second column,

(e) multiply the first row by 5 and add the result to the fourth row,

(f) delete the third row and divide the first row by (—1) and the second row by 2.

We have

$$A \sim \begin{bmatrix} 2 & -4 & 5 & 3 \\ 3 & -6 & 4 & 2 \\ 4 & -8 & 17 & 11 \end{bmatrix} \overset{(a)}{\sim} \begin{bmatrix} 2 & 5 & 3 \\ 3 & 4 & 2 \\ 4 & 17 & 11 \end{bmatrix} \overset{(b)}{\sim} \begin{bmatrix} -5 & -1 & 3 \\ 0 & 0 & 2 \\ -25 & -5 & 11 \end{bmatrix}$$

$$\overset{(c)}{\sim} \begin{bmatrix} -1 & 3 \\ 0 & 2 \\ -5 & 11 \end{bmatrix} \overset{(d)}{\sim} \begin{bmatrix} -1 & 0 \\ 0 & 2 \\ -5 & -4 \end{bmatrix} \overset{(e)}{\sim} \begin{bmatrix} -1 & 0 \\ 0 & 2 \\ 0 & -4 \end{bmatrix}$$

$$\overset{(f)}{\sim} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Since $r(A) = 2$, i.e. $r \leqslant \min\{m, n\}$, the given system has a fundamental system of $n - 2 = 4 - 2 = 2$ solutions.

(2) We shall seek the general solution of the system. We find the base minor, i.e. a second-order nonzero minor. Such is, for instance, the minor composed of the coefficients of x_3 and x_4 in the first and the second equation of the system: $M_2 = \begin{vmatrix} 5 & 3 \\ 4 & 2 \end{vmatrix} = 2 \neq 0$. Leaving the base unknowns x_3 and x_4 on the left-hand side and transferring the constant unknowns x_1 and x_2 to the right-hand side, we arrive at a system

$$\begin{cases} 5x_3+3x_4=-2x_1+4x_2, \\ 4x_3+2x_4=-3x_1+6x_2. \end{cases}$$

Its solution, found from Cramer's formulas, has the form

$$x_3 = -2.5x_1 + 5x_2,$$
$$x_4 = \ \ 3.5x_1 + 7x_2.$$

(3) To obtain a fundamental system of solutions, we must find any two linearly independent solutions of the system (since $n - r = 2$). Setting first $x_1 = 1$, $x_2 = 0$, we have $x_3 = -2.5$, $x_4 = 3.5$; setting then $x_1 = 0$, $x_2 = 1$, we get $x_3 = 5$, $x_4 = -7$. Thus the fundamental system of solutions has the form

$$R_1 = \begin{pmatrix} 1 \\ 0 \\ -2.5 \\ 3.5 \end{pmatrix}, \quad R_2 = \begin{pmatrix} 0 \\ 1 \\ 5 \\ -7 \end{pmatrix},$$

and the general solution is $R = c_1R_1 + c_2R_2$, where c_1 and c_2 are arbitrary numbers.

Assigning different values to c_1 and c_2, we can get any solution of the given system.

Assume, for instance, that $x_1 = 1$, $x_2 = 2$. Then $x_3 = -2.5 \cdot 1 + 5 \cdot 2 = 7.5$, $x_4 = 3.5 \cdot 1 - 7 \cdot 2 = -10.5$. The special solution

$$R = \begin{pmatrix} 1 \\ 2 \\ 7.5 \\ -10.5 \end{pmatrix}$$

obtained is a linear combination of the solutions which constitute a fundamental system for $c_1 = 1$, $c_2 = 2$, $R = R_1 + 2R_2$. ▲

3.6. Basic Elimination Procedure

Consider a system of n linear equations in n unknowns

$$\left\{\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1q}x_q + \ldots + a_{1n}x_n = a_{1,n+1}, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2q}x_q + \ldots + a_{2n}x_n = a_{2,n+1}, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{p1}x_1 + a_{p2}x_2 + \ldots + a_{pq}x_q + \ldots + a_{pn}x_n = a_{p,n+1}, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nq}x_q + \ldots + a_{nn}x_n = a_{n,n+1}. \end{array}\right. \tag{1}$$

We set up an augmented matrix of system (1):

$$\overline{A} = \begin{bmatrix} a_{11} a_{12} a_{13} & \ldots & a_{1q} & \ldots & a_{1n} a_{1,n+1} \\ a_{21} a_{22} a_{23} & \ldots & a_{2q} & \ldots & a_{2n} a_{2,n+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{p1} a_{p2} a_{p3} & \ldots & \boxed{a_{pq}} & \ldots & a_{pn} a_{p,n+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} a_{n2} a_{n3} & \ldots & a_{nq} & \ldots & a_{nn} a_{n,n+1} \end{bmatrix}.$$

We choose the largest in absolute value nonzero element a_{pq} of matrix $\bar{A}$ which does not belong to the column of constant terms. This element is known as the *pivot element*. We calculate the multipliers $m_i = -a_{iq}/a_{pq}$ for all rows with the numbers $i \neq p$ (the pth row which contains the pivot element is known as the *pivot row*).

Then we add, to every nonpivot ith row, the pivot row multiplied by the corresponding multiplier m_i for that row. For example,

$$b_{11}^{(1)} = a_{11} - a_{p1}\frac{a_{1q}}{a_{pq}},$$

$$b_{n2}^{(1)} = a_{n2} - a_{p2}\frac{a_{nq}}{a_{pq}} \text{ and so on.}$$

We obtain a new matrix, in which all the elements of the qth column, except for a_{pq}, consist of zeros:

$$a_{1q} - a_{pq}\frac{a_{1q}}{a_{pq}} = 0,$$

$$a_{2q} - a_{pq}\frac{a_{2q}}{a_{pq}} = 0,$$

$$\cdots\cdots\cdots\cdots$$

$$a_{nq} - a_{pq}\frac{a_{nq}}{a_{pq}} = 0.$$

Deleting this column and the pivot pth row, we get a new matrix $B^{(1)}$, the number of whose rows and columns is smaller by unity. We repeat the operations for the matrix $B^{(1)}$ and get a matrix $B^{(2)}$ and so on.

We thus construct a sequence of matrices $\bar{A}$, $B^{(1)}$, $B^{(2)}$, . . ., $B^{(n-1)}$, the last of which is a binomial row matrix (the pivot row). To find the unknowns x_i, we combine all the pivot rows, beginning with the last row, into a system.

This method of solving a system of linear equations in n unknowns is the **basic elimination procedure** (also known as the method of pivot selection and as pivotal-condensation method). The necessary condition for its application is that $\det A \neq 0$.

Example 1. Use the method of pivot selection to solve the system

$$\begin{cases} 3x_1+2x_2+x_3=5, \\ 2x_1+5x_2+x_3=-3, \\ 2x_1+x_2+3x_3=11. \end{cases}$$

△ (1) We compose an augmented matrix $\bar{A}$ of this system, find the pivot element and calculate the multipliers m_i:

$$\bar{A}=\begin{bmatrix} 3 & 2 & 1 & 5 \\ 2 & \boxed{5} & 1 & -3 \\ 2 & 1 & 3 & 11 \end{bmatrix},$$

$a_{22} = 5$ is the pivot element, $m_1 = -2/5$, $m_3 = -1/5$.

(2) We seek the matrix $B^{(1)}$. We have

$$b_{11}^{(1)}=3-2\cdot\frac{2}{5}=\frac{11}{5}, \quad b_{12}^{(1)}=1-1\cdot\frac{2}{5}=\frac{3}{5},$$

$$b_{13}^{(1)}=5+3\cdot\frac{2}{5}=\frac{31}{5}, \quad b_{21}^{(1)}=2-2\cdot\frac{1}{5}=\frac{8}{5},$$

$$b_{22}^{(1)}=3-1\cdot\frac{1}{5}=\frac{14}{5}, \quad b_{23}^{(1)}=11+3\cdot\frac{1}{5}=\frac{58}{5},$$

$$B^{(1)}=\begin{bmatrix} 1/5 & 3/5 & 31/5 \\ 8/5 & \boxed{14/5} & 58/5 \end{bmatrix}.$$

Here $b_{22}^{(1)} = 14/5$ is the pivot element, $m_1 = -1/14$.

(3) We seek $B^{(2)}$. We have

$$b_{11}^{(2)}=\frac{11}{5}-\frac{8}{5}\cdot\frac{3}{14}=\frac{13}{7}, \quad b_{12}^{(2)}=\frac{31}{5}-\frac{58}{5}\cdot\frac{3}{14}=\frac{26}{7},$$

$$B^{(2)}=[13/7 \;\; 26/7].$$

(4) Using the pivot rows, we arrive at a system

$$\begin{cases} 2x_1+5x_2+x_3=-3, \\ \quad x_1+(14/5)\,x_3=58/5, \\ \quad x_1=26/7 \end{cases} \quad \text{or} \quad \begin{cases} 2x_1+5x_2+x_3=-3, \\ 4x_1+7x_3=29 \\ 13x_1=26. \end{cases}$$

The reverse operation yields $x_1 = 2$, $x_3 = 3$, $x_2 = -2$. ▲

To solve systems of linear equations by the method of pivot selection, we can use the scheme presented in Table 3.1 (the pivot elements in it, which have been chosen arbitrarily, are framed and the pivot rows are labelled by roman figures I-IV).

Table 3.1

m_i	Columns of the system matrix				Constant terms	Σ	Pivot rows	Matrices
	x_1	x_2	x_3	x_4				
m_1	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	$\sum a_{1j}$	II	$\overline{A}$
	a_{21}	a_{22}	$\boxed{a_{23}}$	a_{24}	a_{25}	$\sum a_{2j}$		
m_3	a_{31}	a_{32}	a_{33}	a_{34}	a_{35}	$\sum a_{3j}$		
m_4	a_{41}	a_{42}	a_{43}	a_{44}	a_{45}	$\sum a_{4j}$		
	$b_{11}^{(1)}$	$\boxed{b_{12}^{(1)}}$		$b_{14}^{(1)}$	$b_{15}^{(1)}$	$\sum b_{1j}^{(1)}$	I	$B^{(1)}$
$m_3^{(1)}$	$b_{31}^{(1)}$	$b_{32}^{(1)}$		$b_{34}^{(1)}$	$b_{35}^{(1)}$	$\sum b_{3j}^{(1)}$		
$m_4^{(1)}$	$b_{41}^{(1)}$	$b_{42}^{(1)}$		$b_{44}^{(1)}$	$b_{45}^{(1)}$	$\sum b_{4j}^{(1)}$		
	$\boxed{b_{31}^{(2)}}$			$b_{34}^{(2)}$	$b_{35}^{(2)}$	$\sum b_{3j}^{(2)}$	III	$B^{(2)}$
$m_4^{(2)}$	$b_{41}^{(2)}$			$b_{44}^{(2)}$	$b_{45}^{(2)}$	$\sum b_{4j}^{(2)}$		
				$\boxed{b_{44}^{(3)}}$	$b_{45}^{(3)}$	$\sum b_{4j}^{(3)}$	IV	$B^{(3)}$

Example 2. Use the method of pivot selection to solve the system

$$\begin{cases} x_1+2x_2+3x_3+4x_4=5, \\ 2x_1+x_2+2x_3+3x_4=1, \\ 3x_1+2x_2+x_3+2x_4=1, \\ 4x_1+3x_2+2x_3+x_4=-5. \end{cases}$$

△ The solution is given in Table 3.2.
Thus we obtain a system

$$\begin{cases} -x_3=3, \\ -(4/3)\,x_2-(2/3)\,x_3=-2/3, \\ (15/4)\,x_1+(5/2)\,x_2+(5/4)\,x_3=-25/4, \\ x_1+2x_2+3x_3+4x_4=5, \end{cases}$$

whence we find that $x_3=-3$, $x_2=2$, $x_1=-2$, $x_4=3$. ▲

Table 3.2

m_i	Columns of the system matrix				Constant terms	Σ	Pivot rows	Matrices
	x_1	x_2	x_3	x_4				
	1	2	3	$\boxed{4}$	5	15	I	$\overline{A}$
−3/4	2	1	2	3	1	9		
−1/2	3	2	1	2	1	9		
−1/4	4	3	2	1	−5	5		
−1/3	5/4	−1/2	−1/4		−11/4	−9/4	III	$B^{(1)}$
−2/3	5/2	1	−1/2		−3/2	1		
	$\boxed{15/4}$	5/2	5/4		−25/4	5/4		
−1/2		$\boxed{-4/3}$	−2/3		−2/3	−8/3	I	$B^{(2)}$
		−2/3	−4/3		8/3	2/3		
			−1		3	2	IV	$B^{(3)}$

3.7. Solving Systems of Linear Equations by the Gauss Elimination Method

The **Gaussian elimination** is the most popular method of solving systems of linear equations. It is a special case of the basic elimination method when the upper left nonzero element of the system matrix being considered is chosen as the pivot element.

Here is an example of solving a system of four equations in four unknowns by this method. Consider a system

$$\begin{cases} a_{11}x_1+a_{12}x_2+a_{13}x_3+a_{14}x_4=a_{15}, \\ a_{21}x_1+a_{22}x_2+a_{23}x_3+a_{24}x_4=a_{25}, \\ a_{31}x_1+a_{32}x_2+a_{33}x_3+a_{34}x_4=a_{35}, \\ a_{41}x_1+a_{42}x_2+a_{43}x_3+a_{44}x_4=a_{45}. \end{cases} \quad (1)$$

We shall eliminate the unknown x_1 from all equations of system (1), except for the first equation. We call x_1 the *pivot unknown* and the coefficient a_{11}, the *pivot coefficient.* Dividing the first equation by a_{11} (this is possible if $a_{11} \neq 0$), we get

$$x_1 + \frac{a_{12}}{a_{11}} x_2 + \frac{a_{13}}{a_{11}} x_3 + \frac{a_{14}}{a_{11}} x_4 = \frac{a_{15}}{a_{11}} .$$

We designate $a_{12}/a_{11} = b_{12}$, $a_{13}/a_{11} = b_{13}$, $a_{14}/a_{11} = b_{14}$, $a_{15}/a_{11} = b_{15}$ and, in general, $b_{ij} = a_{1j}/a_{11}$ ($j > 1$). Then the equation being considered assumes the form

$$x_1 + b_{12}x_2 + b_{13}x_3 + b_{14}x_4 = b_{15}, \qquad (2)$$

or

$$x_1 = b_{15} - b_{12}x_2 - b_{13}x_3 - b_{14}x_4.$$

To eliminate the unknown x_1 from the equations of system (1), we perform the following transformations.

(1) We subtract equation (2) multiplied by a_{21} from the second equation of system (1),

$$\begin{array}{l} a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 = a_{25} \\ \underline{-a_{21}x_1 - a_{21}b_{12}x_2 - a_{21}b_{13}x_3 - a_{21}b_{14}x_4 = -a_{21}b_{15}} \\ (a_{22} - a_{21}b_{12})\,x_2 + (a_{23} - a_{21}b_{13})\,x_3 + (a_{24} - a_{21}b_{14})\,x_4 = (a_{25} - a_{21}b_{15}) \end{array},$$

designate

$$a_{22} - a_{21}b_{12} = a_{22}^{(1)};\quad a_{23} - a_{21}b_{13} = a_{23}^{(1)},$$
$$a_{24} - a_{21}b_{14} = a_{24}^{(1)};\quad a_{25} - a_{21}b_{15} = a_{25}^{(1)}$$

and rewrite the resulting equation in the form

$$a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + a_{24}^{(1)}x_4 = a_{25}^{(1)}.$$

(2) From the third equation of system (1) we subtract equation (2) multiplied by a_{31}:

$$\begin{array}{l} a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 = a_{35} \\ \underline{-a_{31}x_1 - a_{31}b_{12}x_2 - a_{31}b_{13}x_3 - a_{31}b_{14}x_4 = -a_{31}b_{15}} \\ (a_{32} - a_{31}b_{12})\,x_2 + (a_{33} - a_{31}b_{13})\,x_3 + (a_{34} - a_{31}b_{14})\,x_4 = a_{35} - a_{31}b_{15} \end{array}.$$

Designating $a_{32} - a_{31}b_{12} = a_{32}^{(1)}$, $a_{33} - a_{31}b_{13} = a_{33}^{(1)}$ and so on, we rewrite the resulting equation in the form

$$a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 + a_{34}^{(1)}x_4 = a_{35}^{(1)}.$$

(3) From the fourth equation of system (1) we subtract equation (2) multiplied by a_{41}. Using similar designations,

we get the following equation:

$$a^{(1)}_{42}x_2 + a^{(1)}_{43}x_3 + a^{(1)}_{44}x_4 = a^{(1)}_{45}.$$

As a result of these elementary transformations, we have a system of three equations in three unknowns

$$\begin{cases} a^{(1)}_{22}x_2 + a^{(1)}_{23}x_3 + a^{(1)}_{24}x_4 = a^{(1)}_{25}, \\ a^{(1)}_{32}x_2 + a^{(1)}_{33}x_3 + a^{(1)}_{34}x_4 = a^{(1)}_{35}, \\ a^{(1)}_{42}x_2 + a^{(1)}_{43}x_3 + a^{(1)}_{44}x_4 = a^{(1)}_{45}, \end{cases} \qquad (1')$$

where the coefficients a_{ij} $(i,\ j \geqslant 2)$ can be found from the formula $a^{(1)}_{ij} = a_{ij} - a_{i1}b_{1j}$ (say, $a^{(1)}_{23} = a_{23} - a_{21}b_{12}$).

Dividing then the coefficients of the first equation of system (1′) by the pivot coefficient $a^{(1)}_{22} \neq 0$, we get the first equation of the system in the form

$$x_2 + \frac{a^{(1)}_{23}}{a^{(1)}_{22}}x_3 + \frac{a^{(1)}_{24}}{a^{(1)}_{22}}x_4 = \frac{a^{(1)}_{25}}{a^{(1)}_{22}}$$

We designate

$$a^{(1)}_{23}/a^{(1)}_{22} = b^{(1)}_{23},\quad a^{(1)}_{24}/a^{(1)}_{22} = b^{(1)}_{24},\quad a^{(1)}_{25}/a^{(1)}_{22} = b^{(1)}_{25}$$

and, in general, $a^{(1)}_{2j}/a^{(1)}_{22} = b^{(1)}_{2j}$ $(j > 2)$. Then the first equation of system (1′) assumes the form

$$x_2 + b^{(2)}_{23}x_3 + b^{(1)}_{24}x_4 = b^{(1)}_{25}, \qquad (2')$$

or

$$x_2 = b^{(1)}_{25} - b^{(1)}_{23}x_3 - b^{(1)}_{24}x_4.$$

Eliminating now x_2 from all equations of system (1′), except for the first, in the same way as we eliminated x_1, we arrive at the following system of two equations in two unknowns:

$$\begin{cases} a^{(2)}_{33}x_3 + a^{(2)}_{34}x_4 = a^{(2)}_{35}, \\ a^{(2)}_{43}x_3 + a^{(2)}_{44}x_4 = a^{(2)}_{45}, \end{cases} \qquad (1'')$$

where $a^{(2)}_{ij} = a^{(1)}_{ij} - a^{(1)}_{i2}\, b^{(1)}_{2j}$ $(i,\ j \geqslant 3)$. (For example, $a^{(2)}_{34} = a^{(1)}_{34} - a^{(1)}_{32}\, b^{(1)}_{24}$.) Dividing the coefficients of the first equation of system (1″) by the pivot coefficient $a^{(2)}_{33} \neq 0$, we get

$$x_3 + b^{(2)}_{34}x_4 = b^{(2)}_{35}, \qquad (2'')$$

where $b_{3j}^{(2)} = a_{3j}^{(2)}/a_{33}^{(2)}$ ($j > 3$), i.e.

$$x_3 = b_{35}^{(2)} - b_{34}^{(2)}x_4.$$

Eliminating now x_3 similarly from system (1″), we find that

$$a_{44}^{(3)}x_4 = a_{45}^{(3)}, \tag{1‴}$$

where $a_{ij}^{(3)} = a_{ij}^{(2)} - a_{i3}^{(2)}b_{3j}^{(2)}$ ($i, j \geqslant 4$). Hence

$$x_4 = a_{45}^{(3)}/a_{44}^{(3)}. \tag{2‴}$$

The other unknowns of the system can be found in succession from equations (2″), (2′) and (2):

$$x_3 = b_{35}^{(2)} - b_{34}^{(2)}x_4,$$

$$x_2 = b_{25}^{(1)} - b_{24}^{(1)}x_4 - b_{23}^{(1)}x_3,$$

$$x_1 = b_{15} - b_{14}x_4 - b_{13}x_3 - b_{12}x_2.$$

Thus the Gaussian elimination reduces to constructing an equivalent system of equations (2), (2′), (2″), (2‴). The Gaussian elimination can be used when all the pivot coefficients are different from zero.

For the sake of convenience, we carry out the calculations according to a *scheme of unique division*. The calculation of the elements b_{ij} is a *forward substitution* and the calculation of the values of the unknowns is a *back substitution* since we first determine the value of the last unknown.

The scheme for the unique division (Gauss' scheme) is composed as follows.

Section I of the scheme (see Table 3.3) includes the coefficients of the unknowns (in the columns of the corresponding unknowns), constant terms and, for each row, "control sums" (column $\sum_2$), equal to the sum of the elements a_{ij} in that row (here $i = 1, 2, 3, 4$, $j = 1, 2, 3, 4, 5$); the last row of Section I consisting of 1 and the elements b_{ij} results from the division of the first row of that section by the pivot coefficient a_{11}.

The elements of Section II of the scheme are equal to the corresponding elements of Section I minus the product $a_{i1}b_{1j}$ ($i, j \geqslant 2$); for instance $a_{23}^{(1)} = a_{23} - a_{21}b_{13}$. The last row of Section II consisting of 1 and the elements

x_1	x_2	x_3	x_4	Constant terms
$\boxed{a_{11}}$ a_{21} a_{31} a_{41}	a_{12} a_{22} a_{32} a_{42}	a_{13} a_{23} a_{33} a_{43}	a_{14} a_{24} a_{34} a_{44}	a_{15} a_{25} a_{35} a_{45}
$1=\frac{a_{11}}{a_{11}}$	$b_{12}=\frac{a_{12}}{a_{11}}$	$b_{13}=\frac{a_{13}}{a_{11}}$	$b_{14}=\frac{a_{14}}{a_{11}}$	$b_{15}=\frac{a_{15}}{a_{11}}$
	$\boxed{a_{22}^{(1)}}$ $a_{32}^{(1)}$ $a_{42}^{(1)}$	$a_{23}^{(1)}$ $a_{33}^{(1)}$ $a_{43}^{(1)}$	$a_{24}^{(1)}$ $a_{34}^{(1)}$ $a_{44}^{(1)}$	$a_{25}^{(1)}=a_{25}-a_{21}b_{15}$ $a_{35}^{(1)}=a_{35}-a_{31}b_{15}$ $a_{45}^{(1)}=a_{45}-a_{41}b_{15}$
	$1=\frac{a_{22}^{(1)}}{a_{22}^{(1)}}$	$b_{23}^{(1)}=\frac{a_{23}^{(1)}}{a_{22}^{(1)}}$	$b_{24}^{(1)}=\frac{a_{24}^{(1)}}{a_{22}^{(1)}}$	$b_{25}^{(1)}=\frac{a_{25}^{(1)}}{a_{22}^{(1)}}$
		$\boxed{a_{33}^{(2)}}$ $a_{43}^{(2)}$	$a_{34}^{(2)}$ $a_{44}^{(2)}$	$a_{35}^{(2)}=a_{35}^{(1)}-a_{32}^{(1)}b_{25}^{(1)}$ $a_{45}^{(2)}=a_{45}^{(1)}-a_{42}^{(1)}b_{25}^{(1)}$
		$1=\frac{a_{33}^{(2)}}{a_{33}^{(2)}}$	$b_{34}^{(2)}=\frac{a_{34}^{(2)}}{a_{33}^{(2)}}$	$b_{35}^{(2)}=\frac{a_{35}^{(2)}}{a_{33}^{(2)}}$
			$\boxed{a_{44}^{(3)}}$	$a_{45}^{(3)}=a_{45}^{(2)}-a_{43}^{(2)}b_{35}^{(2)}$
			$1=\frac{a_{44}^{(3)}}{a_{44}^{(3)}}$	$b_{45}^{(3)}=\frac{a_{45}^{(3)}}{a_{44}^{(3)}}$
1	1	1	1	$x_4=b_{45}^{(3)}$ $x_3=b_{35}^{(2)}-b_{34}^{(2)}x_4$ $x_2=b_{25}^{(1)}-b_{24}^{(1)}x_4-b_{23}^{(1)}x_3$ $x_1=b_{15}-b_{14}x_4-b_{13}x_3$ $-b_{12}x_2$

Table 3.3

Σ_1	Σ_2	Sections of the scheme	
	$a_{16}=a_{11}+a_{12}+a_{13}+a_{14}+a_{15}$ $a_{26}=a_{21}+a_{22}+a_{23}+a_{24}+a_{25}$ $a_{36}=a_{31}+a_{32}+a_{33}+a_{34}+a_{35}$ $a_{46}=a_{41}+a_{42}+a_{43}+a_{44}+a_{45}$	I	Forward substitution
$b_{16}=\frac{b_{16}}{a_{11}}$	$b_{16}=1+b_{12}+b_{13}+b_{14}+b_{15}$		
$a_{26}^{(1)}=a_{26}-a_{21}b_{16}$ $a_{36}^{(1)}=a_{36}-a_{31}b_{16}$ $a_{46}^{(1)}=a_{46}-a_{41}b_{16}$	$a_{26}^{(1)}=a_{22}^{(1)}+a_{23}^{(1)}+a_{24}^{(1)}+a_{25}^{(1)}$ $a_{36}^{(1)}=a_{32}^{(1)}+a_{33}^{(1)}+a_{34}^{(1)}+a_{35}^{(1)}$ $a_{46}^{(1)}=a_{42}^{(1)}+a_{43}^{(1)}+a_{44}^{(1)}+a_{45}^{(1)}$	II	
$b_{26}^{(1)}=\frac{a_{26}^{(1)}}{a_{22}^{(1)}}$	$b_{26}^{(1)}=1+b_{23}^{(1)}+b_{24}^{(1)}+b_{25}^{(1)}$		
$a_{36}^{(2)}=a_{36}^{(1)}-a_{32}^{(1)}b_{26}^{(1)}$ $a_{46}^{(2)}=a_{46}^{(1)}-a_{42}^{(1)}b_{26}^{(1)}$	$a_{36}^{(2)}=a_{33}^{(2)}+a_{34}^{(2)}+a_{35}^{(2)}$ $a_{46}^{(2)}=a_{43}^{(2)}+a_{44}^{(2)}+a_{45}^{(2)}$	III	
$b_{36}^{(2)}=\frac{a_{36}^{(2)}}{a_{33}^{(2)}}$	$b_{36}^{(2)}=1+b_{34}^{(2)}+b_{35}^{(2)}$		
$a_{46}^{(3)}=a_{46}^{(2)}-a_{43}^{(2)}b_{36}^{(2)}$	$a_{46}^{(3)}=a_{44}^{(3)}+a_{45}^{(3)}$	IV	
$b_{46}^{(3)}=\frac{a_{46}^{(3)}}{a_{44}^{(3)}}$	$b_{46}^{(3)}=1+b_{45}^{(3)}$		
$\bar{x}_4=b_{46}^{(3)}$ $\bar{x}_3=b_{36}^{(2)}-b_{34}^{(2)}\bar{x}_4$ $\bar{x}_2=b_{26}^{(1)}-b_{24}^{(1)}\bar{x}_4-b_{23}^{(1)}\bar{x}_3$ $\bar{x}_1=b_{16}-b_{14}\bar{x}_4$ $-b_{13}\bar{x}_3-b_{12}\bar{x}_2$	$\bar{x}_4=1+x_4$ $\bar{x}_3=1+x_3$ $\bar{x}_2=1+x_2$ $\bar{x}_1=1+x_1$	V	Back substitution

$b_{2j}^{(1)}$ results from the division of the first row of that section by the pivot coefficient $a_{22}^{(1)}$.

The elements of the third and fourth sections of the scheme are calculated in a similar way. Sections I, II, III and IV, which end with the calculation of the elements $b_{ij}^{(i-1)}$ ($i = 1, 2, 3, 4$, $j = 2, 3, 4, 5$) constitute a **forward substitution** of the calculation of the scheme.

The **back substitution** begins with the calculation of the last unknown x_4 of the system of linear equations and ends with the calculation of the first unknown x_1. In the back substitution only the rows of the forward substitution are used which contain unities and the corresponding elements b_{ij} (we call these rows "marked").

The element $b_{45}^{(3)}$ of the last "marked" row and the column of constant terms yields the value of x_4. Then the other unknowns x_3, x_2 and x_1 are found by subtracting the sum of the products of its coefficients by the corresponding values of the unknowns found before, say, $x_3 = b_{35}^{(2)} - b_{34}^{(2)} x_4$ from the constant term of the "marked" row.

The values of the unknowns are written in succession in Sec. V. The unities written there allow us to find the corresponding coefficients for x_j in the "marked" rows.

The so-called control sums which are in the column $\sum_2$ are used to verify the calculations

$$a_{i6}=\sum_{j=1}^{5} a_{ij} \ (i = 1, 2, 3, 4) \text{ and } b_{i6}=\sum_{j=1}^{5} b_{ij}+1 \ (i = 1, 2, 3, 4).$$

In the column $\sum_1$ of Sections II, III and IV the same actions are performed on the control sums in each row as on the other elements of that row. If there are no errors in calculations, the elements of the columns $\sum_1$ and $\sum_2$ are equal. Thus we control the forward substitution of the scheme.

To control the back substitution, we find $\bar{x}_4$ in the last "marked" row of the column $\sum_1$, i.e. $\bar{x}_4 = b_{46}^{(3)}$ and find the other unknowns of this column $\bar{x}_j$ ($j = 3, 2, 1$) in the same rows and from the same formulas as the unknowns x_j, with the only difference that we substitute the appropriate $\bar{x}_j$ into the formulas. In the result, the numbers $\bar{x}_j$ must coincide with the numbers $x_j + 1$ of the column $\sum_2$.

Table 3.4

x_1	x_2	x_3	x_4	Constant terms	Σ_1	Σ_2	Sections of the scheme
$\boxed{2}$	2	-1	1	4	8		I
4	3	-1	2	6	14		
8	5	-3	4	12	26		
3	3	-2	2	6	12		
1	1	-0.5	0.5	2	4	4	
	$\boxed{-1}$	1	0	-2	-2	-2	II
	-3	1	0	-4	-6	-6	
	0	-0.5	0.5	0	0	0	
	1	-1	0	2	2	2	
		$\boxed{-2}$	0	2	0	0	III
		-0.5	0.5	0	0	0	
		1	0	-1	0	0	
			$\boxed{0.5}$	-0.5	0	0	IV
			1	-1	0	0	
			1	$x_4=-1$	$\bar{x}_4=0$	0	V
		1		$x_3=-1$	$\bar{x}_3=0$	0	
	1			$x_2=1$	$\bar{x}_2=2$	2	
1				$x_1=1$	$\bar{x}_1=2$	2	

Example 1. Use the scheme of unique division to solve the following system:

$$\begin{cases} 2x_1+2x_2-x_3+x_4=4, \\ 4x_1+3x_2-x_3+2x_4=6, \\ 8x_1+5x_2-3x_3+4x_4=12, \\ 3x_1+3x_2-2x_3+2x_4=6. \end{cases}$$

△ In section I of Table 3.4 we write the matrix of the system, its constant terms and control sums. Then we calculate the "marked" row of this section dividing its first row by $a_{11} = 2$, for instance $b_{12} = a_{12}/a_{11} = 2/2 = 1$.

The elements of Section II are calculated according to the following rule: every element of this section is equal to the corresponding element of Section I minus the product of the first element of its row by the element of the "marked" row in its column. We write the result in the respective place in Section II. For example,

$$a_{23}^{(1)} = a_{23} - a_{21}b_{13} = -1 - 4\,(-0.5) = 1,$$
$$a_{33}^{(1)} = a_{33} - a_{31}b_{13} = -3 - 8\,(-0.5) = 1.$$

We obtain the elements of the "marked "row of section II by dividing its first row by the pivot coefficient $a_{22}^{(1)} = -1$, for instance,

$$b_{23}^{(1)} = a_{23}^{(1)}/a_{22}^{(1)} = 1/(-1) = -1.$$

By analogy we calculate the elements of sections III and IV, for example,

$$a_{44}^{(2)} = a_{44}^{(1)} - a_{42}^{(1)}b_{24}^{(1)} = 2 - 3\cdot 0.5 = 0.5,$$
$$a_{45}^{(3)} = a_{45}^{(2)} - a_{43}^{(2)}b_{35}^{(2)} = 0 - (-0.5)\,(-1) = -0.5.$$

To calculate the elements of section V, i.e. to find the unknowns, we use the "marked" rows, beginning with the last row.

The unknown x_4 is a constant term of the last "marked" row: $x_4 = b_{45}^{(3)} = 1$, and the other unknowns x_3, x_2 and x_1 result from the successive subtraction, from the constant terms of the "marked" rows, of the sums of the products of the corresponding coefficients $b_{ij}^{(i-1)}$ by the values of the unknowns obtained before.

The verification is done with the aid of columns $\sum_1$ and $\sum_2$. The same actions are performed on the column $\sum_1$ as on the other columns (see Tables 3.3 and 3.4) and, as a result, the sum of the elements of every row of the scheme (without the column $\sum_1$) must be equal to the element of that row from the column $\sum_2$. The numbers $\bar{x}_j$ from the column $\sum_1$ must be equal to the numbers $1 + \bar{x}_j$ from the column $\sum_2$.

As a result we obtain $x_1 = 1$, $x_2 = 1$, $x_3 = -1$, $x_4 = -1$. ▲

Example 2. Using the scheme of unique division, solve the following system with an accuracy of 0.0001:

$$\begin{cases} 0.63x_1 + 1.00x_2 + 0.71x_3 + 0.34x_4 = 2.08, \\ 1.17x_1 + 0.18x_2 - 0.65x_3 + 0.71x_4 = 0.17, \\ 2.71x_1 - 0.75x_2 + 1.17x_3 - 2.35x_4 = 1.28, \\ 3.58x_1 + 0.28x_2 - 3.45x_3 - 1.18x_4 = 0.05. \end{cases}$$

△ The solution of the system is given in Table 3.5. The final answer is $x_1 = 0.4026$, $x_2 = 1.5016$, $x_3 = 0.5862$, $x_4 = -0.2678$. ▲

Table 3.5

x_1	x_2	x_3	x_4	Constant terms	Σ_1	Σ_2
$\boxed{0.63}$	1.00	0.71	0.34	2.08	4.76	
1.17	0.18	−0.65	0.71	0.17	1.58	
2.71	−0.75	1.17	−2.35	1.28	2.06	
3.58	0.21	−3.45	−1.18	0.05	−0.79	
1	1.587	1.127	0.539	3.302	7.555	
	$\boxed{-1.6768}$	−1.9686	0.0794	−3.6933	−7.2593	−7.2593
	−5.0508	−1.8842	−3.8107	−7.6684	−18.4141	−18.4141
	−5.4715	−7.4847	−3.1096	−11.7712	−27.8370	−27.8370
	1	1.17402	−0.04735	2.20259	4.32926	4.32926
		$\boxed{4.04554}$	−4.04986	3.45644	3.45212	3.45212
		−1.06105	−3.36868	0.28027	−4.14946	−4.14946
		1	−1.00106	0.85438	0.85332	0.85332
			$\boxed{-4.43085}$	1.18681	−3.24404	−3.24404
			1	−0.26785	0.73215	0.73215
			1	$x_4 = -0.26785$	$\bar{x}_4 = 0.73215$	0.73215
		1		$x_3 = 0.58625$	$\bar{x}_3 = 1.58625$	1.58625
	1			$x_2 = 1.50164$	$\bar{x}_2 = 2.50164$	2.50164
1				$x_1 = 0.40257$	$\bar{x}_1 = 1.40257$	1.40257

If the approximate values of the unknowns obtained according to Gauss' scheme are sufficiently accurate, i.e. the corrections are small in absolute value, the refinement is not necessary.

When it is necessary to make the approximate values of the unknowns more accurate, we must do the following:

(1) for every equation of the system we calculate the errors, i.e. the differences between the right-hand and left-hand sides of the system resulting from the substitution of the approximate values of the unknowns into the equations; if we designate the approximate values of the unknowns as $x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}$, the errors as $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ and the constant terms as $b_1, b_2, \ldots, b_n$, then

$$\varepsilon_1 = b_1 - \sum_{j=1}^{n} a_{ij} x_j^{(0)},$$

$$\varepsilon_2 = b_2 - \sum_{j=1}^{n} a_{2j} x_j^{(0)},$$

$$\varepsilon_n = b_n - \sum_{j=1} a_{nj} x_j^{(0)};$$

(2) we write the errors ε_i $(i = 1, 2, \ldots, n)$ in a separate column ε of Gauss' scheme and perform the same actions on it as on the other columns of the scheme,

(3) considering the column ε to be a column of constant terms, we obtain the corrections δ_i for the unknowns,

(4) we find the precise values of the unknowns adding the corrections δ_i to the corresponding approximate values of the unknowns $x_i^{(0)}$:

$$x_1 = x_1^{(0)} + \delta_1, \ x_2 = x_2^{(0)} + \delta_2, \ldots, x_n = x_n^{(0)} + \delta_n.$$

Example 3. Use Gauss' method accurate to three decimal places to solve the system

$$\begin{cases} 7.09x_1 + 1.17x_2 - 2.23x_3 = -4.75, \\ 0.43x_1 + 1.4x_2 - 0.62x_3 = -1.05, \\ 3.21x_1 - 4.25x_2 + 2.13x_3 = 5.06 \end{cases}$$

and make the approximate values of the unknowns obtained accurate to 10^{-4}.

△ Using Gauss' scheme, we calculate $x_1^{(0)}$, $x_2^{(0)}$ and $x_3^{(0)}$ with three significant digits (Table 3.6).

Table 3.6

x_1	x_2	x_3	Constant terms	Σ	ε
$\boxed{7.09}$	1.17	−2.23	−4.75	1.28	0.00097
0.43	1.4	−0.62	−1.05	0.16	0.00087
3.21	−4.25	2.13	5.06	6.15	−0.00295
1	0.1650	−0.3145	−0.6700	0.1805	0.00014
	$\boxed{1.3290}$	−0.4847	−0.7619	0.0824	0.00081
	−4.7706	3.1395	7.2107	5.5706	−0.00340
	1	−0.3647	−0.5733	0.0620	0.00061
		$\boxed{1.3964}$	4.4705	5.8669	−0.00048
		1	3.2015	4.2015	−0.00035
		1	3.2015	4.2015	−0.00035
	1		0.5943	1.5943	−0.00048
1			0.2388	1.2388	−0.0005

Thus $x_1^{(0)}=0.239$, $x_2^{(0)}=0.594$, $x_3^{(0)}=3.202$.

To find the correction $\delta=\begin{bmatrix}\delta_1\\ \delta_2\\ \delta_3\end{bmatrix}$, we must solve the given system with the same matrix

$$A=\begin{bmatrix}7.09 & 1.17 & -2.23\\ 0.43 & 1.4 & -0.62\\ 3.21 & -4.25 & 2.13\end{bmatrix}$$

and new constant terms ε_i (errors) which we shall calculate as follows.

1. We calculate the errors, for which purpose we substitute the values of $x_1^{(0)}$, $x_2^{(0)}$, $x_3^{(0)}$ into the equations of the system:

$$7.09\cdot 0.239 + 1.17\cdot 0.594 - 2.23\cdot 3.202 = -4.75097,$$
$$0.43\cdot 0.239 + 1.4\cdot 0.594 - 0.62\cdot 3.202 = -1.05087,$$
$$3.21\cdot 0.239 - 4.25\cdot 0.594 - 2.13\cdot 3.202 = 5.06295.$$

The errors are equal, respectively, to

$$\varepsilon_1 = -4.75 - (-4.75097) = 0.00097,$$
$$\varepsilon_2 = -1.05 - (-1.05087) = 0.00087,$$
$$\varepsilon_3 = 5.06 - 5.06295 = -0.00295.$$

2. To solve the given system, we use Gauss' scheme with constant terms $\varepsilon_1 = 0.00097$, $\varepsilon_2 = 0.00087$, $\varepsilon_3 = -0.00295$. Correspondingly, with an accuracy of 10^{-4}, we get the values of the corrections $\delta_1 = -0.0004$, $\delta_2 = 0.0005$, $\delta_3 = -0.0001$. Then we make the unknowns more precise:

$$x_1 = x_1^{(0)} + \delta_1 = 0.239 - 0.0004 = 0.2386;$$
$$x_2 = x_2^{(0)} + \delta_2 = 0.594 + 0.0005 = 0.5945;$$
$$x_3 = x_3^{(0)} + \delta_3 = 3.202 - 0.0001 = 3.2019. \blacktriangle$$

3.8. Calculating Determinants by the Gauss Elimination Method

Gaussian elimination can be used to calculate the determinants:

$$\begin{vmatrix} a_{11} & a_{12} & a_{1n} \\ a_{21} & a_{22} & a_{2n} \\ & \cdot & \\ a_{n1} & a_{n2} & a_{nn} \end{vmatrix} = a_{11}a_{22}^{(1)}a_{33}^{(2)} \ldots a_{nn}^{(n-1)},$$

where a_{11}, a_{22}, $a_{33}^{(2)}$, ..., $a_{nn}^{(n-1)}$ are the pivot elements of the scheme of the unique division.

Example 1. Using the scheme of unique division, calculate the determinant

$$d = \begin{vmatrix} 1 & 1 & 2 & 3 \\ 3 & -1 & -1 & -2 \\ 2 & 3 & -1 & -1 \\ 1 & 2 & 3 & -1 \end{vmatrix}.$$

$\triangle$ The solution is given in Table 3.7.

Table 3.7

Columns				Σ_1	Σ_2
1	2	3	4		
$\boxed{1}$	1	2	3	7	
3	−1	−1	−2	−1	
2	3	−1	−1	3	
1	2	3	−1	5	
1	1	2	3	7	7
	$\boxed{-4}$	−7	−11	−22	−22
	1	−5	−7	−11	−11
	1	1	−4	−2	−2
	1	7/4	11/4	22/4	22/4
		$\boxed{-27/4}$	−39/4	−66/4	−66/4
		−3/4	−27/4	−30/4	−30/4
		1	13/9	22/9	22/9
			$\boxed{-17/3}$	−17/3	−17/3

Thus $d = 1\cdot(-4)\cdot(-27/4)\cdot(-17/3) = -153$. ▲

Example 2. Using the scheme of unique division, calculate the determinant

$$d = \begin{vmatrix} 1.00 & 0.42 & 0.54 & 0.66 \\ 0.42 & 1.00 & 0.32 & 0.44 \\ 0.54 & 0.32 & 1.00 & 0.22 \\ 0.66 & 0.44 & 0.22 & 1.00 \end{vmatrix}$$

with an accuracy of 0.001.

△ The solution is given in Table 3.8.

The final result is $d = 1\cdot 0.8236\cdot 0.6978\cdot 0.4979 = 0.286$. ▲

Table 3.8

Columns				Σ_1	Σ_2
1	2	3	4		
$\boxed{1}$	0.42	0.54	0.66	2.62	
0.42	1.00	0.32	0.44	2.18	
0.54	0.32	1.00	0.22	2.08	
0.66	0.44	0.22	1.00	2.32	
1	0.42	0.54	0.66	2.62	
	$\boxed{0.8236}$	0.0932	0.1628	1.0796	1.0796
	0.0932	0.7084	0.1364	0.6652	0.6652
	0.1628	−0.1364	0.5644	0.5908	0.5908
	1	0.1135	0.1973	1.3108	1.3108
		$\boxed{0.6978}$	−0.1548	0.5430	0.5430
		−0.1549	0.5323	0.3774	0.3774
		1	−0.2219	0.7782	0.7781
			$\boxed{0.4979}$	0.4979	0.4979

3.9. The Gaussian Elimination for Inversion of a Matrix

Consider a nonsingular matrix $A = [a_{ij}]$ $(i, j = 1, 2, \ldots, n)$. To find its inverse $A^{-1} = [x_{ij}]$, we use the fundamental relation $AA^{-1} = I$, where I is an nth-order identity matrix.

Thus, for a fourth-order matrix, performing the multiplication

$$AA^{-1} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \cdot \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

we get four systems of equations for 16 unknowns x_{ij} $(i, j = 1, 2, 3, 4)$.

In the general case, we have relations

$$\sum_{k=1}^{n} a_{ik} \cdot x_{kj} = \delta_{ij} \quad (i, j = 1, 2, \ldots, n),$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{for } i \neq j, \end{cases}$$

δ_{ij} is the *Kronecker delta*.

Then n systems of linear equations obtained for $j = 1, 2, \ldots, n$ have the same matrix A and different constant terms which constitute an identity matrix, and therefore we can use Gaussian elimination method to solve these systems.

The solutions x_{ij} found according to the scheme of unique division are the elements of the inverse matrix A^{-1}.

Example 1. Use Gauss' method to invert the matrix

$$A = \begin{bmatrix} 1 & 0 & 1 & 2 \\ -1 & 2 & 3 & 1 \\ 4 & 0 & -2 & 1 \\ 0 & 2 & 1 & 2 \end{bmatrix}.$$

△ The solution is given in Table 3.9.
Thus

$$A^{-1} = \begin{bmatrix} 0 & 1/3 & 1/3 & -1/3 \\ -1/2 & 1/6 & 1/6 & 1/3 \\ 1/5 & 7/15 & 1/15 & -7/15 \\ 2/5 & -2/5 & -1/5 & 2/5 \end{bmatrix}. \blacktriangle$$

Example 2. Invert the matrix

$$A = \begin{bmatrix} 1.00 & 0.47 & -0.11 & 0.55 \\ 0.42 & 1.00 & 0.35 & 0.17 \\ -0.25 & 0.67 & 1.00 & 0.36 \\ 0.54 & -0.32 & -0.74 & 1.00 \end{bmatrix}$$

using Gaussian elimination method. All the calculations must be done to four decimal places. Round off the answer to three decimal places.

△ The solution is given in Table 3.10.

Table 3.9

x_{1j}	x_{2j}	x_{3j}	x_{4j}	$j=1$	$j=2$	$j=3$	$j=4$	Σ_1	Σ_2
$\boxed{1}$	0	1	2	1	0	0	0	5	
−1	2	3	1	0	1	0	0	0	
4	0	−2	1	0	0	1	0	4	
0	2	1	2	0	0	0	1	6	
1	0	1	2	1	0	0	0	5	5
	$\boxed{2}$	4	3	1	1	0	0	11	11
	0	−6	−7	−4	0	1	0	−16	−16
	2	1	2	0	0	0	1	6	6
	1	2	3/2	1/2	1/2	0	0	11/2	11/2
		$\boxed{-6}$	−7	−4	0	1	0	−16	−16
		3	−7	−1	−1	0	1	−5	−5
		1	7/6	4/6	0	−1/6	0	16/6	16/6
			$\boxed{5/2}$	1	−1	−1/2	1	3	3
			1	2/5	−2/5	−1/5	2/5	6/5	6/5
			1	2/5	−2/5	−1/5	2/5	6/5	6/5
		1		1/5	7/15	1/15	−7/15	19/15	19/15
	1			−1/2	1/6	1/6	1/3	7/6	7/6
1				0	1/3	1/3	−1/3	4/3	4/3

Consequently,

$$A^{-1}=\begin{bmatrix} 1.9759 & -1.2017 & -0.0120 & -0.8781 \\ -1.2883 & 2.1003 & -0.4869 & 0.5268 \\ 1.4921 & -1.7239 & 1.0873 & -0.9189 \\ -0.3750 & 0.0453 & 0.6553 & 0.9626 \end{bmatrix}$$

$$\approx\begin{bmatrix} 1.976 & -1.202 & -0.012 & -0.878 \\ -1.288 & 2.100 & -0.487 & 0.527 \\ 1.492 & -1.724 & 1.087 & -0.919 \\ -0.375 & 0.045 & 0.655 & 0.963 \end{bmatrix}. \blacktriangle$$

Table 3.10

x_{1j}	x_{2j}	x_{3j}	x_{4j}	$j = 1$
$\boxed{1.00}$	0.47	−0.11	0.55	1
0.42	1.00	0.35	0.17	0
−0.25	0.67	1.00	0.36	0
0.54	−0.32	−0.74	1.00	0
1	0.47	−0.11	0.55	1
	$\boxed{0.8026}$	0.3962	−0.0610	−0.4200
	0.7875	0.9725	0.4975	0.2500
	−0.5738	−0.6806	0.7030	−0.5400
	1	0.4936	−0.0760	−0.5233
		$\boxed{0.5838}$	0.5573	0.6621
		−0.3974	0.6594	−0.8403
		1	0.9546	1.1341
			$\boxed{1.0388}$	−0.3896
			1	−0.3750
1	1	1	1	$x_{41} = -0.3750$ $x_{31} = 1.4921$ $x_{21} = -1.2883$ $x_{11} = 1.9759$

$j = 2$	$j = 3$	$j = 4$	Σ_1	Σ_2
0	0	0	2.91	
1	0	0	2.94	
0	1	0	2.78	
0	0	1	1.48	
0	0	0	2.91	2.91

Table 3.10 continued

$j=2$	$j=3$	$j=4$	Σ_1	Σ_2
1	0	0	1.7178	1.7178
0	1	0	3.5075	3.5075
0	0	1	−0.0914	−0.0914
1.2460	0	0	2.1403	2.1403
−0.9812	1	0	1.8220	1.8220
0.7150	0	1	1.1367	1.1367
−1.6807	1.7129	0	3.1209	3.1209
0.0471	0.6807	1	2.3770	2.3770
0.0453	0.6553	0.9626	2.2882	2.2882
$x_{42}=0.0453$	$x_{43}=0.6553$	$x_{44}=0.9626$	2.2882	2.2882
$x_{32}=-1.7239$	$x_{33}=1.0873$	$x_{34}=-0.9189$	0.9366	0.9366
$x_{22}=2.1003$	$x_{23}=-0.4869$	$x_{24}=0.5268$	1.8519	1.8519
$x_{12}=-1.2017$	$x_{13}=-0.0120$	$x_{14}=-0.8781$	0.8841	0.8841

3.10. Cholesky's Method

Assume that a system of linear equations is given in the matrix form

$$A\mathbf{x} = \mathbf{b}, \tag{1}$$

where $A = [a_{ij}]$ is a square matrix of order n, and

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} a_{1,n+1} \\ a_{2,n+1} \\ \vdots \\ a_{n,n+1} \end{bmatrix}$$ are column vectors.

We represent the matrix A as the product of the lower triangular matrix $C = [c_{ij}]$ and the upper triangular matrix $B = [b_{ij}]$ with a unit diagonal, i.e.

$$A = CB,$$

where

$$C = \begin{bmatrix} c_{11} & 0 & 0 & \dots & 0 \\ c_{21} & c_{22} & 0 & \dots & 0 \\ c_{31} & c_{32} & c_{33} & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ c_{n1} & c_{n2} & c_{n3} & \dots & c_{nn} \end{bmatrix}, \quad B = \begin{bmatrix} 1 & b_{12} & b_{13} & \dots & b_{1n} \\ 0 & 1 & b_{23} & \dots & b_{2n} \\ 0 & 0 & 1 & \dots & b_{3n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}.$$

The elements c_{ij} and b_{ij} are found from formulas (17) and (18) of 2.6:

$$c_{i1} = a_{i1}, \quad c_{ij} = a_{ij} - \sum_{k=1}^{j-1} c_{ik} b_{kj} \quad \text{for } 1 < j \leqslant i, \tag{3}$$

$$b_{1j} = \frac{a_{1j}}{c_{11}}, \quad b_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} c_{ik} b_{kj}}{c_{ii}} \quad \text{for } 1 < i < j. \tag{4}$$

We can write equation (1) in the form

$$CB\mathbf{x} = \mathbf{b}. \tag{5}$$

The product $B\mathbf{x}$ of the matrix B and the column vector $\mathbf{x}$ is a column vector which we designate as $\mathbf{y}$:

$$B\mathbf{x} = \mathbf{y}. \tag{6}$$

Then we can rewrite equation (5) in the form

$$C\mathbf{y} = \mathbf{b}, \tag{7}$$

or

$$\begin{bmatrix} c_{11} & 0 & 0 & \dots & 0 \\ c_{21} & c_{22} & 0 & \dots & 0 \\ c_{31} & c_{32} & c_{33} & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ c_{n1} & c_{n2} & c_{n3} & \dots & c_{nn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} a_{1,n+1} \\ a_{2,n+1} \\ a_{3,n+1} \\ \vdots \\ a_{n,n+1} \end{bmatrix}. \tag{7'}$$

The elements c_{ij} $(i, j = 1, 2, \dots, n)$ here are known since the matrix A of system (1) is assumed to be expanded in the product of two triangular matrices C and B [formulas (3) and (4)].

Multiplying the matrices on the left-hand side of relation (7′), we get a system of equations

$$\left\{\begin{aligned} c_{11}y_1 &= a_{1,\,n+1},\\ c_{21}y_1+c_{22}y_2 &= a_{2,\,n+1},\\ c_{31}y_1+c_{32}y_2+c_{33}y_3 &= a_{3,\,n+1},\\ c_{n1}y_1+c_{n2}y_2+c_{n3}y_3+\ldots+c_{nn}y_n &= a_{n,\,n+1},\end{aligned}\right. \qquad (8)$$

whence we obtain the following formulas for the unknowns:

$$y_1=\frac{a_{1,\,n+1}}{c_{11}}, \quad y_i=\frac{a_{i,\,n+1}-\sum\limits_{k=1}^{i-1} c_{ik}y_k}{c_{ii}}, \quad i>1. \qquad (9)$$

It is convenient to calculate the unknowns y_i together with the elements b_{ij}.

When all y_i ($i = 1, 2, \ldots, n$) are found from formulas (9), we substitute them into equation (6):

$$\begin{bmatrix} 1 & b_{12} & b_{13} & \ldots & b_{1n}\\ 0 & 1 & 0 & \ldots & b_{2n}\\ 0 & 0 & 1 & \ldots & b_{3n}\\ \cdot & \cdot & \cdot & \cdot & \cdot\\ 0 & 0 & 0 & \ldots & 1\end{bmatrix}\begin{bmatrix} x_1\\ x_2\\ x_3\\ \vdots\\ x_n\end{bmatrix}=\begin{bmatrix} y_1\\ y_2\\ y_3\\ \vdots\\ y_n\end{bmatrix}.$$

Multiplying, we get a system

$$\left\{\begin{aligned} x_1+b_{12}x_2+b_{13}x_3+\ldots+b_{1n}x_n &= y_1,\\ x_2+b_{23}x_3+\ldots+b_{2n}x_n &= y_2,\\ x_3+\ldots+b_{3n}x_n &= y_3,\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots&\\ x_n &= y_n.\end{aligned}\right. \qquad (10)$$

Since the coefficients b_{ij} have been determined [see formula (4)], we can calculate the values of the unknowns, beginning with the last one, using the following formulas:

$$x_n=y_n, \quad x_i=y_i-\sum_{k=i+1}^{n} b_{ik}x_k, \quad i<n. \qquad (11)$$

This method is known as **Cholesky's method** (improved elimination method), in which the usual control with the aid of sums is employed.

When we solve systems with the use of Cholesky's method, it is convenient to use Table 3.11 and seek y_i together with the coefficients b_i.

Table 3.11

	x_1	x_2	x_3	x_4	Constant terms	Σ	x_1	x_2	x_3	x_4	Constant terms	Σ
I	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	1	2	−1	2	4	8
	a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a_{26}	2	3	−1	4	6	14
	a_{31}	a_{32}	a_{33}	a_{34}	a_{35}	a_{36}	4	5	−3	8	12	26
	a_{41}	a_{42}	a_{43}	a_{44}	a_{45}	a_{46}	2	3	−2	3	6	12
II	c_{11} 1	b_{12}	b_{13}	b_{14}	$y_1=b_{15}$	b_{16}	1 1	2	−1	2	4	8
	c_{21}	c_{22} 1	b_{23}	b_{24}	$y_2=b_{25}$	b_{26}	2	−1 1	−1	0	2	2
	c_{31}	c_{32}	c_{33} 1	b_{34}	$y_3=b_{35}$	b_{36}	4	−3	−2 1	0	−1	0
	c_{41}	c_{42}	c_{34}	c_{44} 1	$y_4=b_{45}$	b_{46}	2	−1	−1	−1 1	1	2
III	x_1	x_2	x_3	x_4			−1	1	−1	1		

Example. Using Cholesky's method, solve the system

$$\begin{cases} 3x_1-2x_2-5x_3+x_4=3, \\ 2x_1-3x_2+x_3+5x_4=-3, \\ x_1+2x_2 \qquad -4x_4=-3, \\ x_1-x_2-4x_3+9x_4=22. \end{cases}$$

△ The solution is given in Table 3.11.

In the first section of the table we write the matrix of the coefficients, its constant terms and control sums.

Then we fill in section II according to the rule indicated in 2.6, i.e. first find the first column of the matrix C, then the first row of the matrix B, the second column of the matrix C, the second row of the matrix B and so on.

Section III is used to determine x_i.

The verification is done with the aid of the column Σ with which we perform the same operations as with the column of constant terms.

(1) We find the elements of the first column of the matrix C from the formula

$$c_{i1} = a_{i1} \quad (i = 1,\ 2,\ 3,\ 4).$$

Then we write the first column of section I into the first column of section II:

$$c_{11} = a_{11} = 1,\quad c_{21} = a_{21} = 2,\quad c_{31} = a_{31} = 4,\quad c_{41} = a_{41} = 2.$$

(2) We find the elements of the first row of the matrix B from the formula

$$b_{ij} = a_{1j}/c_{11} \quad (j = 2,\ 3,\ 4,\ 5,\ 6),$$

i.e.

$$b_{12} = a_{12}/c_{11} = 2,\quad b_{13} = a_{13}/c_{11} = -1,$$
$$b_{14} = a_{14}/c_{11} = 2,\ y_1 = b_{15} = a_{15}/c_{11} = 4,\ b_{16} = a_{16}/c_{11} = 8,$$
$$b_{16} = 1 + b_{12} + b_{13} + b_{14} + b_{15} = 1 + 2 - 1 + 2 + 4 = 8.$$

(3) We find the elements of the second column of the matrix C from the formula

$$c_{i2} = a_{i2} - c_{i1}b_{12} \quad (i = 2,\ 3,\ 4),$$

i.e.

$$c_{22} = a_{22} - c_{21}b_{12} = 3 - 2\cdot 2 = -1,$$
$$c_{32} = a_{32} - c_{31}b_{12} = 5 - 4\cdot 2 = -3,$$
$$c_{42} = a_{42} - c_{41}b_{12} = 3 - 2\cdot 2 = -1.$$

(4) We find the elements of the second row of the matrix B from the formula

$$b_{2j} = \frac{a_{2j} - c_{21}b_{1j}}{c_{22}} \quad (j = 3,\ 4,\ 5,\ 6),$$

i.e.

$$b_{23}=\frac{a_{23}-c_{21}b_{13}}{c_{22}}=\frac{-1-2\cdot(-1)}{-1}=-1,$$

$$b_{24}=\frac{a_{24}-c_{21}b_{14}}{c_{22}}=\frac{4-2\cdot 2}{-1}=0,$$

$$y_2=b_{25}=\frac{a_{25}-c_{21}b_{15}}{c_{22}}=\frac{6-2\cdot 4}{-1}=2,$$

$$b_{26}=\frac{a_{26}-c_{21}b_{16}}{c_{22}}=\frac{14-2\cdot 8}{-1}=2,$$

$$b_{26}=1+b_{23}+b_{24}+b_{25}=1-1+0+2=2.$$

(5) We find the elements of the third column of the matrix C from the formula

$$c_{i3}=a_{i3}-c_{i1}b_{13}-c_{i2}b_{23}\quad (i=3,\ 4),$$

i.e.

$$c_{33}=a_{33}-c_{31}b_{13}-c_{32}b_{23}=3-4\cdot(-1)-(-3)(-1)=-2,$$
$$c_{43}=a_{43}-c_{41}b_{13}-c_{42}b_{23}=-2-2\cdot(-1)-(-1)\cdot(-1)=-1.$$

(6) We find the elements of the third row of the matrix B from the formula

$$b_{3j}=\frac{a_{3j}-c_{31}b_{1j}-c_{32}b_{2j}}{c_{33}}\qquad (j=4,\ 5,\ 6),$$

$$b_{34}=\frac{a_{34}-c_{31}b_{14}-c_{32}b_{24}}{c_{33}}=\frac{8-4\cdot 2-(-3)\cdot 0}{-2}=0,$$

$$y_3=b_{35}=\frac{a_{35}-c_{31}b_{15}-c_{32}b_{25}}{c_{33}}=\frac{12-4\cdot 4-(-3)\cdot 2}{-2}=-$$

$$b_{36}=\frac{a_{36}-c_{31}b_{16}-c_{32}b_{26}}{c_{33}}=\frac{26-4\cdot 8-(-3)\cdot 2}{-2}=0,$$

$$b_{36}=1+b_{34}+b_{35}=1+0-1=0.$$

(7) We find the elements of the fourth column of the matrix C from the formula

$$c_{44}=a_{44}-c_{41}b_{14}-c_{42}b_{24}-c_{43}b_{34},$$

i.e.

$$c_{44}=3-2\cdot 2-(-1)\cdot 0-4\cdot 0=-1.$$

(8) We find the elements of the fourth row of the matrix B from the formula

$$b_{4j}=\frac{a_{4j}-c_{41}b_{1j}-c_{42}b_{2j}-c_{43}b_{3j}}{c_{44}}\quad (j=5,\ 6),$$

i.e.

$$y_4 = b_{45} = \frac{a_{45} - c_{41}b_{15} - c_{42}b_{25} - c_{43}b_{35}}{c_{44}}$$

$$= \frac{6 - 2\cdot 4 - (-1)\cdot 2 - (-1)(-1)}{-1} = 1,$$

$$b_{46} = \frac{a_{46} - c_{41}b_{16} - c_{42}b_{26} - c_{43}b_{36}}{c_{44}}$$

$$= \frac{12 - 2\cdot 8 - (-1)\cdot 2 - 4\cdot 0}{-1} = 2,$$

$$b_{46} = 1 + b_{45} = 1 + 1 = 2.$$

(9) Then we calculate x_i using the formula

$$x_i = y_i - \sum_{k=i+1}^{n} b_{ik}x_k, \qquad i = 1,\ 2,\ 3,\ 4,$$

where $y_1 = 4$, $y_2 = 2$, $y_3 = -1$ and $y_4 = 1$. We have

$x_4 = y_4 = 1,$

$x_3 = y_3 - b_{34}x_4 = -1 - 0\cdot 1 = -1,$

$x_2 = y_2 - b_{23}x_3 - b_{24}x_4 = 2 - (-1)(-1) - 0\cdot 1 = 1,$

$x_1 = y_1 - b_{12}x_2 - b_{13}x_3 - b_{14}x_4 = 4 - 2\cdot 1 - (-1)(-1)$
$- 2\cdot 1 = -1.$

Thus

$$x_1 = -1, \quad x_2 = 1, \quad x_3 = -1, \quad x_4 = 1. \ ▲$$

3.11. The Iterative Method (the Method of Successive Approximations)

The approximate methods for solving systems of linear equations make it possible to obtain the values of the roots of the system with the specified accuracy as the limit of the sequence of some vectors. The process of constructing such a sequence is known as the *iterative process*.

The efficiency of the application of approximate methods depends on the choice of the initial vector and the rate of convergence of the process.

In this section we consider the **method of iteration** (the **method of successive approximations**). Assume that we are given n linear equations in n unknowns:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2, \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n. \end{cases} \qquad (1)$$

We write system (1) in matrix form:

$$A\mathbf{x} = \mathbf{b}, \tag{2}$$

where

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

Assuming that the diagonal elements $a_{ii} \neq 0$ ($i = 1, 2, \ldots, n$), we express x_1 using the first equation of the system, x_2, using the second equation and so on. As a result we get a system equivalent to system (1):

$$\left\{\begin{aligned} x_1 &= \frac{b_{11}}{a_{11}} - \frac{a_{12}}{a_{11}} x_2 - \frac{a_{13}}{a_{11}} x_3 - \ldots - \frac{a_{1n}}{a_{11}} x_n, \\ x_2 &= \frac{b_2}{a_{22}} - \frac{a_{21}}{a_{22}} x_1 - \frac{a_{23}}{a_{22}} x_3 - \ldots - \frac{a_{2n}}{a_{22}} x_n, \\ &\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot \\ x_n &= \frac{b_n}{a_{nn}} - \frac{a_{n1}}{a_{nn}} x_1 - \frac{a_{n2}}{a_{nn}} x_2 - \ldots - \frac{a_{n,\,n-1}}{a_{nn}} x_{n-1}. \end{aligned}\right. \tag{3}$$

We designate $b_i/a_{ii} = \beta_i$, $-a_{ij}/a_{ii} = \alpha_{ij}$, where $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, n$. Then we can write system (3) as follows:

$$\left\{\begin{aligned} x_1 &= \beta_1 + \alpha_{12}x_2 + \alpha_{13}x_3 + \ldots + \alpha_{1n}x_n, \\ x_2 &= \beta_2 + \alpha_{21}x_1 + \alpha_{23}x_3 + \ldots + \alpha_{2n}x_n, \\ &\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot \\ x_n &= \beta_n + \alpha_{n1}x_1 + \alpha_{n2}x_2 + \ldots + \alpha_{n,\,n-1}x_{n-1}. \end{aligned}\right. \tag{3'}$$

System (3′) is known as a system reduced to the *normal form*. Introducing the designations

$$\alpha = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \alpha_{n1} & \alpha_{n2} & \ldots & \alpha_{nn} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix},$$

we write system (3′) in matrix form:

$$\mathbf{x} = \boldsymbol{\beta} + \alpha\mathbf{x},$$

or

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}. \quad (4)$$

We use the method of successive approximations to solve system (4). For the zeroth approximation we take the column of constant terms:

$$\begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ \vdots \\ x_n^{(0)} \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \text{ is the zeroth approximation.}$$

Then we construct a column matrix

$$\begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \\ \vdots \\ x_n^{(1)} \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nn} \end{bmatrix} \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ \vdots \\ x_n^{(0)} \end{bmatrix} \text{ is the first approximation,}$$

$$\begin{bmatrix} x_1^{(2)} \\ x_2^{(2)} \\ \vdots \\ x_n^{(2)} \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nn} \end{bmatrix} \begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \\ \vdots \\ x_n^{(1)} \end{bmatrix} \text{ is the second approximation,}$$

and so on.

In general, any $(k+1)$th approximation can be calculated by the formula

$$\mathbf{x}^{(k+1)} = \beta + \alpha \mathbf{x}^{(k)} \quad (k = 0, 1, \dots, n). \quad (5)$$

If the sequence of approximation $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(k)}$ has a limit $\mathbf{x} = \lim\limits_{k\to\infty} \mathbf{x}^{(k)}$, then this limit is a solution of system (3) since, by virtue of the property of the limit $\lim\limits_{x\to\infty} \mathbf{x}^{(k+1)} = \beta + \alpha \lim\limits_{k\to\infty} \mathbf{x}^{(k)}$, i.e. $\mathbf{x} = \alpha + \beta\mathbf{x}$.

Example 1. Solve the system

$$\begin{cases} 8x_1 + x_2 + x_3 = 26, \\ x_1 + 5x_2 - x_3 = 7, \\ x_1 - x_2 + 5x_3 = 7 \end{cases}$$

with an accuracy of 10^{-1} using the iterative method.

△ (1) We reduce the system to normal form

$$\begin{cases} x_1 = 3.25 - 0.25x_2 - 0.125x_3, \\ x_2 = 1.4 - 0.2x_1 + 0.2x_3, \\ x_3 = 1.4 - 0.2x_1 + 0.2x_2; \end{cases}$$

$$\alpha = \begin{bmatrix} 0 & -0.125 & -0.125 \\ -0.2 & 0 & 0.2 \\ -0.2 & 0.2 & 0 \end{bmatrix};$$

$$\beta = \begin{bmatrix} 3.25 \\ 1.4 \\ 1.4 \end{bmatrix}.$$

(2) We construct successive approximations. The zeroth approximation is

$$\begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ x_3^{(0)} \end{bmatrix} = \begin{bmatrix} 3.25 \\ 1.4 \\ 1.4 \end{bmatrix}.$$

The first approximation is

$$\begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{bmatrix} = \begin{bmatrix} 3.25 \\ 1.4 \\ 1.4 \end{bmatrix} + \begin{bmatrix} 0 & -0.125 & -0.125 \\ -0.2 & 0 & 0.2 \\ -0.2 & 0.2 & 0 \end{bmatrix}$$

$$\times \begin{bmatrix} 3.25 \\ 1.4 \\ 1.4 \end{bmatrix} = \begin{bmatrix} 2.9 \\ 1.03 \\ 1.03 \end{bmatrix}.$$

The second approximation is

$$\begin{bmatrix} x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{bmatrix} = \begin{bmatrix} 3.25 \\ 1.4 \\ 1.4 \end{bmatrix} + \begin{bmatrix} 0 & -0.125 & -0.125 \\ -0.2 & 0 & 0.2 \\ -0.2 & 0.2 & 0 \end{bmatrix} \begin{bmatrix} 2.9 \\ 1.03 \\ 1.03 \end{bmatrix} = \begin{bmatrix} 2.992 \\ 026 \\ 1.026 \end{bmatrix}$$

The third approximation is

$$\begin{bmatrix} x_1^{(3)} \\ x_2^{(3)} \\ x_3^{(3)} \end{bmatrix} = \begin{bmatrix} 3.25 \\ 1.4 \\ 1.4 \end{bmatrix} + \begin{bmatrix} 0 & -0.125 & -0.125 \\ -0.2 & 0 & 0.2 \\ -0.2 & 0.2 & 0 \end{bmatrix} \begin{bmatrix} 2.992 \\ 1.026 \\ 1.026 \end{bmatrix} = \begin{bmatrix} 2.99 \\ 1.01 \\ 1.01 \end{bmatrix}.$$

Thus $x_1 = 2.99$, $x_2 = 1.01$, $x_3 = 1.01$, and, with an accuracy of 10^{-1}, we obtain $x_1 = 3.0$, $x_2 = 1.0$, $x_3 = 1.0$. ▲

Example 2. Solve the system

$$\begin{cases} 7.6x_1 + 0.5x_2 + 2.4x_3 = 1.9, \\ 2.2x_1 + 9.1x_2 + 4.4x_3 = 9.7, \\ -1.3x_1 + 0.2x_2 + 5.8x_3 = -1.4. \end{cases} \qquad (*)$$

with an accuracy of 10^{-3} using the iterative method.

△ (1) We reduce the system to normal form:

$$\begin{cases} x_1=\dfrac{1.9}{7.6}-\dfrac{0.5}{7.6}x_2-\dfrac{2.4}{7.6}x_3, \\ x_2=\dfrac{9.7}{9.1}-\dfrac{2.2}{9.1}x_1-\dfrac{4.4}{9.1}x_3, \\ x_3=\dfrac{-1.4}{5.8}+\dfrac{1.3}{5.8}x_1-\dfrac{0.2}{5.8}x_2, \end{cases}$$

$$\text{or}\quad \begin{cases} x_1=0.25-0.065x_2-0.3158x_3, \\ x_2=1.0659-0.2418x_1-0.4847x_2, \\ x_3=-0.2414+0.2241x_1-0.3448x_2; \end{cases}$$

$$\alpha=\begin{bmatrix} 0 & -0.065 & -0.3158 \\ -0.2418 & 0 & -0.4847 \\ 0.2241 & -0.3448 & 0 \end{bmatrix}, \quad \beta=\begin{bmatrix} 0.25 \\ -0.0659 \\ -0.2414 \end{bmatrix}$$

Note that we can also do the following to reduce a linear system to normal form: we write the coefficients of x_1, x_2, x_3 in the corresponding equations of system (*) in the form kx, where k is a number close to the coefficient of the respective unknown and by which it is easy to divide the coefficients of the unknowns and constant terms.

For example,

$$\begin{aligned} 10x_1 &= 7.6x_1 + 2.4x_1 \text{ (in the first equation)}, \\ 10x_2 &= 9.1x_2 + 0.9x_2 \text{ (in the second equation)}, \\ 10x_3 &= 5.8x_3 + 4.2x_3 \text{ (in the third equation)}. \end{aligned}$$

We rewrite system (*) as follows:

$$\begin{cases} 10x_1=1.9+2.4x_1-0.5x_2-2.4x_3, \\ 10x_2=9.7-2.2x_1+0.9x_2-4.4x_3, \\ 10x_3=-1.4+1.3x_1-0.2x_2+4.2x_3, \end{cases}$$

$$\begin{cases} x_1=0.19+0.24x_1-0.05x_2-0.24x_3, \\ x_2=0.97-0.22x_1+0.09x_2-0.44x_3, \\ x_3=-0.14+0.13x_1-0.02x_2+0.42x_3. \end{cases}$$

The matrix α and the vector β assume the form

$$\alpha=\begin{bmatrix} 0.24 & -0.05 & -0.24 \\ -0.22 & 0.09 & -0.44 \\ 0.13 & -0.02 & 0.42 \end{bmatrix}, \quad \beta=\begin{bmatrix} 0.19 \\ 0.97 \\ -0.14 \end{bmatrix}$$

(2) We find in succession that

$$\begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ x_3^{(0)} \end{bmatrix}=\begin{bmatrix} 0.19 \\ 0.97 \\ -0.14 \end{bmatrix}.$$

$$\begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{bmatrix} = \begin{bmatrix} 0.19 \\ 0.97 \\ -0.14 \end{bmatrix} + \begin{bmatrix} 0.24 & -0.05 & -0.24 \\ -0.22 & 0.09 & -0.44 \\ 0.13 & -0.02 & 0.42 \end{bmatrix} \times \begin{bmatrix} 0.19 \\ 0.97 \\ -0.14 \end{bmatrix}$$

$$= \begin{bmatrix} 0.2207 \\ 1.0771 \\ -0.1935 \end{bmatrix}.$$

$$\begin{bmatrix} x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{bmatrix} = \begin{bmatrix} 0.19 \\ 0.97 \\ -0.14 \end{bmatrix} + \begin{bmatrix} 0.24 & -0.05 & -0.24 \\ -0.22 & 0.09 & -0.44 \\ 0.13 & -0.02 & 0.42 \end{bmatrix} \times \begin{bmatrix} 0.2207 \\ 1.0771 \\ -0.1935 \end{bmatrix}$$

$$= \begin{bmatrix} 0.2359 \\ 1.1034 \\ -0.2141 \end{bmatrix}.$$

Thus, with an accuracy of 10^{-3}, we obtain

$$x_1 = 0.236,\ x_2 = 1.103,\ x_3 = -0.214. \ \blacktriangle$$

3.12. The Conditions for Convergence of an Iterative Process

Consider a system of linear equations reduced to normal form: $\mathbf{x} = \beta + \alpha x$. The condition for convergence of an iterative process consists in the following: *if the sum of the moduli of the elements of the rows or the sum of the moduli of the elements of the columns, is smaller than unity, then the iterative process for the given system reduces to a unique solution irrespective of the choice of the initial vector.*

Consequently, we can write the condition for convergence as follows:

$$\sum_{j=1}^{n} |\alpha_{ij}| < 1 \quad (i = 1, 2, \ldots, n) \text{ or}$$

$$\sum_{i=1}^{n} |\alpha_{ij}| < 1 \quad (j = 1, 2, \ldots, n).$$

Example. For the system

$$\begin{cases} 8x_1 + x_2 + x_3 = 26, \\ x_1 + 5x_2 - x_3 = 7, \\ x_1 - x_2 + 5x_3 = 7, \end{cases} \text{ or } \begin{cases} x_1 = 3.25 - 0.125x_2 - 0.125x_3, \\ x_2 = 1.4 - 0.2x_1 \quad + 0.2x_3, \\ x_3 = 1.4 - 0.2x_1 \quad + 0.2x_2, \end{cases}$$

the iteration process converges since

$$\alpha = \begin{bmatrix} 0 & 0.125 & -0.125 \\ -0.2 & 0 & 0.2 \\ -0.2 & 0.2 & 0 \end{bmatrix}$$

and

$$|\alpha_{11}|+|\alpha_{21}|+|\alpha_{31}|=0.2+0.2=0.4<1;$$
$$|\alpha_{12}|+|\alpha_{22}|+|\alpha_{32}|=0.125+0.2=0.325<1;$$
$$|\alpha_{13}|+|\alpha_{23}|+|\alpha_{33}|=0.125+0.2=0.325<1.$$

By analogy, we can verify the fulfillment of the condition for convergence taking the sums of the moduli of the elements of the rows.

The iterative process obviously converges if the elements of the matrix α satisfy the inequality $|\alpha_{ij}| < 1/n$, where n is the number of the unknowns of the system. In this example $n = 3$ and all the elements $|\alpha_{ij}| < 1/3$.

The convergence of an iterative process is related to the norms of the matrix α as follows. *If one of the conditions*

$$\|\alpha\|_1 = \max_i \sum_{j=1}^{n} |\alpha_{ij}| < 1$$

or

$$\|\alpha\|_2 = \max_j \sum_{i=1}^{n} |\alpha_{ij}| < 1,$$

or

$$\|\alpha\|_3 = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} |\alpha_{ij}|^2} < 1,$$

is fulfilled, then the process of iteration of the linear system converges to a unique solution.

Thus, in the example considered above, the norm

$$\|\alpha\|_2 = \max(0.4,\ 0.325,\ 0.325) = 0.4 < 1,$$

i.e. the iterative process converges.

3.13. Estimation of the Error of the Approximate Process of the Iterative Method

If the permissible error ε of calculations is specified and $\mathbf{x}_i$ is the vector of the exact values of the unknowns of a linear system and $\mathbf{x}_i^{(k)}$ is the kth approximation of the values of the unknowns calculated with the use of the iterative method, then, to estimate the error $\|\mathbf{x}_i - \mathbf{x}_i^{(k)}\| \leqslant \varepsilon$ of the method, use is made of the formula

$$\|\mathbf{x}_i - \mathbf{x}_i^{(k)}\| \leqslant \frac{\|\alpha\|^{k+1}}{1-\|\alpha\|}\cdot\|\beta\|, \qquad (1)$$

where $\|\alpha\|$ is one of the three norms of the matrix α, $\|\beta\|$ is the same norm of the vector β and k is the number of iterations needed to attain the specified accuracy. In this case we assume that the successive approximations $\mathbf{x}_i^{(j)}$ (where $j = 0, 1, \ldots, k$, $i = 1, 2, \ldots, n$) are calculated exactly, without the round-off errors.

Example. Show that for the system

$$\begin{cases} 9.9x_1 - 1.5x_2 + 2.6x_3 = 0, \\ 0.4x_1 + 13.6x_2 - 4.2x_3 = 8.2, \\ 0.7x_1 + 0.4x_2 + 7.1x_3 = -1.3 \end{cases}$$

the iterative process converges and find out how many iterations must be carried out to find the unknowns with an accuracy of 10^{-4}.

△ (1) We reduce the system to normal form

$$\begin{cases} 10x_1 = 0.1x_1 + 1.5x_2 - 2.6x_3, \\ 20x_2 = -0.4x_1 + 6.4x_2 + 4.2x_3 + 8.2, \\ 10x_3 = -0.7x_1 - 0.4x_2 + 2.9x_3 - 1.3, \end{cases}$$

or

$$\begin{cases} x_1 = 0.01x_1 + 0.15x_2 - 0.26x_3, \\ x_2 = -0.02x_1 + 0.32x_2 + 0.21x_3 + 0.41, \\ x_3 = -0.07x_1 - 0.04x_2 + 0.29x_3 - 0.13. \end{cases}$$

(2) The system matrix

$$\alpha = \begin{bmatrix} 0.01 & 0.15 & -0.26 \\ -0.02 & 0.32 & 0.21 \\ -0.07 & -0.04 & 0.29 \end{bmatrix}.$$

Using the norm $\|\alpha\|_2$ we get $\|\alpha\|_2 = \max(0.1,\ 0.51,\ 0.76) = 0.76 < 1$. Consequently, for this system the iterative process converges.

(3) We have $\beta = \begin{bmatrix} 0 \\ 0.41 \\ -0.13 \end{bmatrix}$, $\|\beta\|_2 = 0 + 0.41 + 0.13 = 0.54$.

(4) Using formula (1), we find that

$$\|\mathbf{x} - \mathbf{x}^{(k)}\| \leqslant \frac{\|\alpha\|_2^{k+1}\cdot\|\beta\|_2}{1-\|\alpha\|_2} = \frac{0.76^{k+1}\cdot 0.54}{0.46} \leqslant 10^{-4},$$

$$0.76^{k+1}\cdot 0.54 \leqslant 10^{-4}\cdot 0.46;\quad 0.76^{k+1} \leqslant \frac{10^{-4}\cdot 46}{54},$$

$$(k+1)\log 0.76 \leqslant \log 46 - \log 54 - 4,$$
$$-(k+1)\cdot 0.1192 \leqslant 1.6628 - 1.7324 - 2 = -4.0696,$$
$$k+1 > \frac{4.0696}{0.1192} = 32.9, \quad k > 32.9, \quad k = 33.$$

The theoretical estimate of the number of iterations necessary to ensure the specified accuracy proves to be too high in practice. ▲

3.14. Seidel's Method. The Conditions for Convergence of Seidel's Process

Seidel's method is a modification of the method of successive approximations. When calculating the $(k+1)$th approximation of the unknown x_i $(i > 1)$ by Seidel's method, we must take into account the k approximations of the unknowns $x_1, x_2, \ldots, x_{i-1}$ found before.

Consider a linear system reduced to normal form

$$\begin{cases} x_1 = \beta_1 + \alpha_{11}x_1 + \alpha_{12}x_2 + \ldots + \alpha_{1n}x_n, \\ x_2 = \beta_2 + \alpha_{21}x_1 + \alpha_{22}x_2 + \ldots + \alpha_{2n}x_n, \\ x_3 = \beta_3 + \alpha_{31}x_1 + \alpha_{32}x_2 + \ldots + \alpha_{3n}x_n, \\ \ldots\ldots\ldots\ldots\ldots \\ x_n = \beta_n + \alpha_{n1}x_1 + \alpha_{n2}x_2 + \ldots + \alpha_{nn}x_n. \end{cases} \tag{1}$$

We arbitrarily choose the initial approximations of the unknowns $x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}$ and substitute them into the first equation of system (1):

$$x_1^{(1)} = \beta_1 + \alpha_{11}x_1^{(0)} + \alpha_{12}x_2^{(0)} + \ldots + \alpha_{1n}x_n^{(0)}.$$

We substitute the first approximation $x_1^{(1)}$ into the second equation of system (1):

$$x_2^{(1)} = \beta_2 + \alpha_{21}x_1^{(1)} + \alpha_{22}x_2^{(0)} + \ldots + \alpha_{2n}x_n^{(0)}.$$

Then we substitute the first approximations $x_1^{(1)}$ and $x_2^{(1)}$ into the third equation of system (1):

$$x_3^{(1)} = \beta_3 + \alpha_{31}x_1^{(1)} + \alpha_{32}x_2^{(1)} + \ldots + \alpha_{3n}x_n^{(0)}$$

and so on. Finally we have

$$x_n^{(1)} = \beta_n + \alpha_{n1}x_1^{(1)} + \alpha_{n2}x_2^{(1)} + \ldots + \alpha_{n,\,n-1}x_{n-1}^{(1)} + \alpha_{nn}x_n^{(0)}$$

By analogy we construct the second, the third, etc. iterations.

Thus, assuming that the kth approximations x_i^k are known, we use Seidel's method to construct the $(k+1)$th approximations by the following formulas:

$$\begin{aligned} x_1^{(k+1)} &= \beta_1 + \sum_{j=1}^{n} \alpha_{1j} x_j^{(k)}, \\ x_2^{(k+1)} &= \beta_2 + \alpha_{21} x_1^{(k+1)} + \sum_{j=2}^{n} \alpha_{2j} x_j^{(k)}, \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ x_n^{(k+1)} &= \beta_n + \sum_{j=1}^{n-1} \alpha_{nj} x_j^{(k+1)} + \alpha_{nn} x_n^{(k)}, \end{aligned} \tag{2}$$

where $k = 0, 1, 2, \ldots, n$.

Example 1. Use Seidel's method to solve the system

$$\begin{cases} 7.6x_1 + 0.5x_2 + 2.4x_3 = 1.9, \\ 2.2x_1 + 9.1x_2 + 4.4x_3 = 9.7, \\ -1.3x_1 + 0.2x_2 + 5.8x_3 = -1.4 \end{cases}$$

with an accuracy of 10^{-3}.

△ (1) We reduce the system to normal form:

$$\begin{cases} 10x_1 = 1.9 + 2.4x_1 - 0.5x_2 - 2.4x_3, \\ 10x_2 = 9.7 - 2.2x_1 + 0.9x_2 - 4.4x_3, \\ 10x_3 = -1.4 + 1.3x_1 - 0.2x_2 + 4.2x_3, \end{cases}$$

or

$$\begin{cases} x_1 = 0.19 + 0.24x_1 - 0.05x_2 - 0.24x_3, \\ x_2 = 0.97 - 0.22x_1 + 0.09x_2 - 0.44x_3, \\ x_3 = -0.14 + 0.13x_1 - 0.02x_2 + 0.42x_3. \end{cases}$$

(2) For the zeroth approximations we take the corresponding values of constant terms: $x_1^{(0)} = 0.19$, $x_2^{(0)} = 0.97$, $x_3^{(0)} = -0.14$.

(3) We construct iterations using Seidel's method. The first approximations are

$$x_1^{(1)} = 0.19 + 0.24 \cdot 0.19 - 0.05 \cdot 0.97 - 0.24 \cdot (-0.14) = 0.2207,$$
$$x_2^{(1)} = 0.97 - 0.22 \cdot 0.2207 + 0.09 \cdot 0.97 - 0.44 \cdot (-0.14) = 1.0703,$$
$$x_3^{(1)} = -0.14 + 0.13 \cdot 0.2207 - 0.02 \cdot 1.0703 + 0.42 \cdot (-0.14) = -0.1915.$$

The second approximations are

$$x_1^{(2)} = 0.19 + 0.24 \cdot 0.2207 - 0.05 \cdot 1.0703 - 0.24 \cdot (-0.1915) = 0.2354,$$
$$x_2^{(2)} = 0.97 - 0.22 \cdot 0.2354 + 0.09 \cdot 1.0703 - 0.44 \cdot (-0.1915)$$

$$= 1.0988,$$

$$x_3^{(2)} = -0.14 + 0.13\cdot 0.2354 - 0.02\cdot 1.0988 + 0.42\cdot(-0.1915)$$

$$= -0.2118$$

and so on.

The solution of this example is given in Table 3.12. The construction of iterations is completed when we get the same values in two successive iterations with the specified degree of accuracy. In this example these are iterations 7 and 8.

The final answer is $x_1 \simeq 0.248$, $x_2 \cong 1.114$, $x_3 \cong -0.224$. ▲

Table 3.12

No. of iteration	x_1	x_2	x_3	No. of iteration	x_1	x_2	x_3
0	0.19	0.97	−0.14	5	0.2467	1.1138	−0.2237
1	0.2207	1.0703	−0.1915	6	0.2472	1.1143	−0.2241
2	0.2354	1.0988	−0.2118	7	0.2474	1.1145	−0.2243
3	0.2424	1.1088	−0.2196	8	0.2475	1.1145	−0.2243
4	0.2454	1.1124	−0.2226				

For the linear system $\mathbf{x} = \beta + \alpha x$ Seidel's process (as well as the process of successive approximations) *converges to a unique solution for any choice of the initial approximation if at least one of the norms of the matrix* α *is smaller than unity, i.e. if*

$$\|\alpha\|_1 = \max_i \sum_{j=1}^{n} |\alpha_{ij}| < 1$$

or

$$\|\alpha\|_2 = \max_j \sum_{i=1}^{n} |\alpha_{ij}| < 1,$$

or

$$\|\alpha\|_3 = \sqrt{\sum_{i,j} |\alpha_{ij}|^2} < 1.$$

Seidel's process converges to a unique solution faster than the process of a simple iteration.

Example 2. Verify whether Seidel's process converges for the system considered in Example 1.

△ (1) After reducing the system to normal form (see p. 157), we get a matrix

$$\alpha=\begin{bmatrix} 0.24 & -0.05 & -0.24 \\ -0.22 & 0.09 & 0.44 \\ 0.13 & -0.02 & 0.42 \end{bmatrix}.$$

(2) We find that $\|\alpha\|_1=\max\limits_i \sum\limits_j |\alpha_{ij}|=\max(0.53, 0.75, 0.57)=0.75<1$. Consequently, for this system the iteration process converges to a unique solution despite the fact that $\|\alpha\|_2=\max\limits_j \sum\limits_{i=1}^{n} |\alpha_{ij}|=\max(0.59, 0.16, 1.1)=1.1>1$. ▲

3.15. Estimation of the Errors of Seidel's Process

Consider a linear system $\mathbf{x}=\beta+\alpha\mathbf{x}$. If $\mathbf{x}_i$ is the exact value of the unknowns of the linear system and $\mathbf{x}_i^{(k)}$ is the kth approximation calculated by Seidel's method, then, to calculate the error of this method, use is made of the formula

$$\|\mathbf{x}-\mathbf{x}^{(k)}\|_1\leqslant\frac{\|\alpha\|_1^{(k)}}{1-\|\alpha\|_1}\|\mathbf{x}^{(1)}-\mathbf{x}^{(0)}\|_1. \qquad (1)$$

Example. Find how many iterations Seidel's method requires to find a solution of the system

$$\begin{cases} 9.9x_1-1.5x_2+2.6x_3=0, \\ 0.4x_1+13.6x_2-4.2x_3=8.2, \\ 0.7x_1+0.4x_2+7.1x_3=-1.3, \end{cases}$$

with an accuracy of 10^{-4}.

△ (1) We reduce the system to normal form (see p. 155):

$$\begin{cases} x_1=0.01x_1+0.15x_2-0.26x_3, \\ x_2=0.41-0.02x_1+0.32x_2+0.21x_3, \\ x_3=-0.13-0.07x_1-0.04x_2+0.29x_3. \end{cases}$$

(2) For the zeroth approximations we take the column of constant terms $x_1^{(0)}=0$, $x_2^{(0)}=0.41$, $x_3^{(0)}=-0.13$ and calculate the first approximations:

$$x_1^{(1)}=0.01\cdot0+0.15\cdot0.41-0.26\cdot(-0.13)=0.0953,$$

$$x_2^{(1)}=0.41-0.02\cdot0.0953+0.32\cdot0.41+0.21\cdot(-0.13)=0.5120,$$

$$x_3^{(1)}=-0.13-0.07\cdot0.0953-0.04\cdot0.5120+0.29\cdot(-0.13)$$
$$=-0.1948.$$

(3) The matrix

$$\alpha=\begin{bmatrix} 0.01 & 0.15 & -0.26 \\ -0.02 & 0.32 & 0.21 \\ -0.07 & -0.04 & 0.29 \end{bmatrix}.$$

This means that $\| \alpha \|_1 = \max(0.42, 0.55, 0.40) = 0.55$. Since

$$\mathbf{x}^{(0)}=\begin{bmatrix} 0 \\ 0.41 \\ -0.13 \end{bmatrix} \quad \text{and} \quad \mathbf{x}^{(1)}=\begin{bmatrix} 0.0953 \\ 0.5120 \\ -0.1948 \end{bmatrix},$$

we have

$$\mathbf{x}^{(1)}-\mathbf{x}^{(0)}=\begin{bmatrix} 0.0953 \\ 0.1120 \\ -0.0648 \end{bmatrix}, \text{ i.e. } \|\mathbf{x}^{(1)}-\mathbf{x}^{(0)}\|_1=0.1120.$$

(4) We find k from formula (1):

$$10^{-4} \leqslant \frac{0.55^k}{0.45} \cdot 0.1120, \quad 10^{-4} \cdot 0.45 \leqslant 0.55^k \cdot 0.1120$$

$$-4\log 10+\log 0.45 \leqslant k\log 0.55+\log 0.1120,$$

$$-4-0.3468 \leqslant k(-0.2596-0.9508), \quad k \geqslant \frac{4.3468}{1.2104}=3.59, \quad k=4.$$

By analogy we can estimate Seidel's method using the norm $\| \alpha \|_2$. ▲

3.16. Reducing a System of Linear Equations to a Form Convenient for Iterations

For the linear system $\mathbf{x} = \beta + \alpha\mathbf{x}$ the process of successive approximations and Seidel's process converge to a unique solution irrespective of the choice of the initial vector if

$$\sum_{j=1}^{n} |\alpha_{ij}| < 1 \quad (i=1, 2, \ldots, n) \text{ or}$$

$$\sum_{i=1}^{n} |\alpha_{ij}| < 1 \quad (j=1, 2, \ldots, n).$$

Thus, for these iterative processes to converge, it is sufficient that the values of the elements α_{ij} of the matrix α, for $i \neq j$, be not very large in absolute value. This is equivalent to the fact that if, for the linear system $A\mathbf{x} = \mathbf{b}$, the moduli of the diagonal coefficients of each equation of the system are larger than the sum of the moduli of all other coefficients (not counting the constant terms), then the iterative processes converge for this system, i.e. if we have a system $\sum_{j=1}^{n} a_{ij}x_j = b_i$ $(i = 1, 2, \ldots, n)$, with $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$, then the process of successive approximations and Seidel's process converge for this system.

Using elementary transformations, we can replace the linear system $A\mathbf{x} = \mathbf{b}$ by the equivalent system $\mathbf{x} = \boldsymbol{\beta} + \boldsymbol{\alpha}\mathbf{x}$ for which the conditions of convergence are fulfilled.

Example. Reduce the system of linear equations

$$\begin{cases} 0.9x_1+2.7x_2-3.8x_3=2.4, & \text{(A)} \\ 2.5x_1+5.8x_2-0.5x_3=3.5, & \text{(B)} \\ 4.5x_1-2.1x_2+3.2x_3=-1.2 & \text{(C)} \end{cases}$$

to a form convenient for iterations.

△ (1) From this system we isolate the equations with coefficients whose moduli are larger than the sum of the moduli of the other coefficients of the system. We write every isolated equation as a row of the new system so that the coefficient, which is the largest in absolute value, would be diagonal.

In equation (B) the coefficient of x_2 is larger in absolute value than the sum of the moduli of the other coefficients. We assume equation (B) to be the second equation of the new system:

$$2.5x_1 + 5.8x_2 - 0.5x_3 = 3.5. \qquad \text{(II)}$$

(2) From the remaining equations of the system we compose linearly independent combinations. Thus, we can take the linear combination (2C) + (A) as the first equation of the system, and then we have

$$9.9x_1 - 1.5x_2 + 2.6x_3 = 0. \qquad \text{(I)}$$

As the third equation of the new system we can take the linear combination (2A) — (B), i.e.

$$0.7x_1 + 0.4x_2 + 7.1x_3 = -1.3. \qquad \text{(III)}$$

(3) As a result we obtain a transformed system of linear equations (I), (II), (III) which is equivalent to the initial system and satisfies the conditions of convergence of the iteration process:

$$\begin{cases} 9.9x_1-1.5x_2+2.6x_3=0, \\ 2.5x_1+5.8x_2-0.5x_3=3.5, \\ 0.7x_1+0.4x_2+7.1x_3=-1.3. \end{cases}$$

Reducing this system to normal form, we have

$$\begin{cases} x_1 = 0.1x_1 + 0.15x_2 - 0.26x_3, \\ x_2 = 0.35x_1 - 0.21x_2 + 0.42x_3 + 0.05 \\ x_3 = -0.13x_1 - 0.07x_2 - 0.04x_3 + 0.29, \end{cases}$$

$$\alpha = \begin{bmatrix} 0.10 & 0.15 & -0.26 \\ 0.35 & -0.21 & 0.42 \\ -0.13 & -0.07 & -0.04 \end{bmatrix};$$

$$\| \alpha \|_2 = \max (0.58;\ 0.43;\ 0.72) = 0.72 < 1.$$

It remains to solve the system using one of the iterative methods. ▲

Exercises

1. Solve the following homogeneous systems of equations:

$$\text{(a)} \begin{cases} x_1 + 3x_2 + 2x_3 = 0, \\ 2x_1 - x_2 + 3x_3 = 0, \\ 3x_1 - 5x_2 + 4x_3 = 0, \\ x_1 + 17x_2 + 4x_3 = 0; \end{cases} \quad \text{(b)} \begin{cases} 3x_1 + 4x_2 - 5x_3 + 7x_4 = 0, \\ 2x_1 - 3x_2 + 3x_3 - 2x_4 = 0, \\ 4x_1 + 11x_2 - 13x_3 + 16x_4 = 0, \\ 7x_1 - 2x_2 + x_3 + 3x_4 = 0. \end{cases}$$

2. Test the following systems of equations for consistence and find its general solution and one special solution:

$$\text{(a)} \begin{cases} 2x_1 + 7x_2 + 3x_3 + x_4 = 6, \\ 3x_1 + 5x_2 + 2x_3 + 2x_4 = 4, \\ 9x_1 + 4x_2 + x_3 + 7x_4 = 2; \end{cases} \quad \text{(b)} \begin{cases} 2x_1 - 3x_2 + 5x_3 + 7x_4 = 1, \\ 4x_1 - 6x_2 + 2x_3 + 3x_4 = 2, \\ 2x_1 - 3x_2 - 11x_3 - 15x_4 = 1. \end{cases}$$

3. Use Cramer's formulas to solve the following systems of linear equations:

$$\text{(a)} \begin{cases} 2x_1 + x_2 + 4x_3 = 7, \\ 2x_1 - x_2 - x_3 = -5, \\ 3x_1 + 4x_2 - 5x_3 = -14; \end{cases} \quad \text{(b)} \begin{cases} 11x + 3y - z = 15, \\ 2x + 5y + 5z = -11, \\ x + y + z = 1. \end{cases}$$

4. Use Gaussian elimination method to solve the following systems:

$$\text{(a)} \begin{cases} x_1 - 4x_2 - x_4 = 6, \\ x_1 + x_2 + 2x_3 + 3x_4 = -1, \\ 2x_1 + 3x_2 - x_3 - x_4 = -1, \\ x_1 + 2x_2 + 3x_3 - x_4 = 3; \end{cases} \quad \text{(b)} \begin{cases} 2x_1 - x_3 - 2x_4 = -8, \\ x_2 + 2x_3 - x_4 = -1, \\ x_1 - x_2 - x_4 = -6, \\ -x_1 + 3x_2 - 2x_3 = 7. \end{cases}$$

5. Use Gaussian elimination method to solve the following systems with an accuracy of 0.001:

$$\text{(a)} \begin{cases} 1.14x_1 - 2.15x_2 - 5.11x_3 = 2.05, \\ 0.42x_1 - 1.13x_2 + 7.05x_3 = 0.80, \\ -0.71x_1 + 0.81x_2 - 0.02x_3 = -1.07, \end{cases}$$

$$\text{(b)} \begin{cases} 0.61x + 0.71y - 0.05z = -0.16 \\ -1.03x - 2.05y + 0.87z = 0.50, \\ 2.5x - 3.12y + 5.03z = 0.95. \end{cases}$$

6. Use Gaussian elimination method to calculate the determinants:

$$\text{(a) } d=\begin{vmatrix} 1 & 4 & 1 & 3 \\ 0 & -1 & 3 & -1 \\ 3 & 1 & 0 & 2 \\ 1 & -2 & 5 & 1 \end{vmatrix}, \quad \text{(b) } d=\begin{vmatrix} -1.6 & 5.4 & -7.7 & 3.1 \\ 8.2 & 1.4 & -2.3 & 0.2 \\ 5.3 & -5.9 & 2.7 & -7.9 \\ 0.7 & 1.9 & -8.5 & 4.8 \end{vmatrix}$$

7. Use Gaussian elimination method to invert the following matrices:

$$\text{(a) } A=\begin{bmatrix} 1 & 2 & 2 & -1 \\ 2 & 7 & 6 & -1 \\ 0 & 3 & 1 & 4 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \quad \text{(b) } A=\begin{bmatrix} 0.32 & 0.52 & -0.42 & 0.23 \\ 0.44 & -0.25 & 0.36 & -0.51 \\ -1.06 & 0.74 & -0.83 & 0.48 \\ 0.96 & 0.82 & 0.55 & 0.36 \end{bmatrix}$$

Carry out the calculations to three decimal places and round off the answer to two decimal digits.

8. Use the method of successive approximations to solve the following systems of linear equations with an accuracy of 0.01, first determining the necessary number of iterations:

$$\text{(a) } \begin{cases} 8.7x_1-3.1x_2+1.8x_3-2.2x_4=-9.7, \\ 2.1x_1+6.7x_2-2.2x_3 = 13.1, \\ 3.2x_1-1.8x_2-9.5x_3-1.9x_4=6.9, \\ 1.2x_1+2.8x_2-1.4x_3-9.9x_4=25.1, \end{cases}$$

$$\text{(b) } \begin{cases} 6.1x+0.7y-0.05z=6.97, \\ -1.3x-2.05y+0.87z=0.10, \\ 2.5x-3.12y-5.03z=2.04. \end{cases}$$

9. Use Seidel's method to solve the systems of linear equations given in Exercise 8, first determining the necessary number of iterations.

10. For the matrices

$$\text{(a) } A=\begin{bmatrix} -0.3 & 1.2 & -0.2 \\ -0.1 & -0.2 & 1.6 \\ -1.5 & -0.3 & 0.1 \end{bmatrix}, \quad \text{(b) } A=\begin{bmatrix} 0.2 & 0.44 & 0.81 \\ 0.58 & -0.29 & 0.05 \\ 0.05 & 0.34 & 0.1 \end{bmatrix}$$

calculate the norms $\|A\|_1$, $\|A\|_2$ and $\|A\|_3$.

11. Use the method of pivotal condensation to solve the following systems:

$$\text{(a) } \begin{aligned} 3x_1-2x_2-5x_3+x_4&=-5, \\ 2x_1-3x_2+x_3+5x_4&=7, \\ x_1+2x_2-4x_4&=-1, \\ x_1-x_2-3x_3+9x_4&=-4; \end{aligned}$$

$$\text{(b) } \begin{aligned} 4x_1-3x_2+x_3-5x_4&=7, \\ 7x_1-2x_2-3x_3-2x_4&=6, \\ 3x_1-2x_2+5x_3-2x_4&=0, \\ 2x_1+3x_2+5x_3+4x_4&=-5. \end{aligned}$$

12. Use Cholesky's scheme to solve the following systems of linear equations:

a) $$\begin{cases} x_1+2x_2+3x_3+4x_4=5, \\ 2x_1+x_2+2x_3+3x_4=1, \\ 3x_1+2x_2+x_3+2x_4=1, \\ 4x_1+3x_2+2x_3+x_4=-5; \end{cases}$$

(b) $$\begin{cases} x_1+2x_2+3x_3=0, \\ 2x_1+x_2+2x_3=1, \\ 3x_1+2x_2+x_3=4. \end{cases}$$

Chapter 4

Calculating the Values of Elementary Functions

In calculations we often encounter the problem of finding the values of a function at a specified point. In that case we must bear in mind that mathematically equivalent expressions not always prove to be of the same value when they are calculated. For example, to calculate the left-hand side of the identity

$$a^2 + 2ab + b^2 = (a + b)^2,$$

we have to perform four operations of multiplication and two operations of addition, and to calculate its right-hand side, we have to make only one addition and one multiplication.

Thus we arrive at a significant problem of representing a function in optimal form in order to construct the algorithm of calculation of its values. We can understand the optimality in the sense of minimization of either the total number of arithmetic operations or the time needed to calculate the values of the function.

4.1. Calculating the Values of Algebraic Polynomials

An *algebraic polynomial* of degree n is an expression of the form

$$P(x) = a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n, \qquad (1)$$

where the coefficients $a_0, a_1, \ldots, a_n$ are real numbers and a_0 is assumed to be nonzero. The coefficient a_n is a *constant term* of polynomial (1).

The calculation of the values of an algebraic polynomial is a typical example of minimization of the number of computing operations. This problem is important in practical applications not only in its direct sense but also because it is closely connected with the problem of division of a polynomial by a linear binomial (a first-degree polynomial), i.e. with the problem of finding the roots of a polynomial.

A *root* (a *zero*) *of polynomial* (1) is every number ξ which turns this polynomial into zero when ξ is substituted for x in the polynomial, i.e. $P(\xi) = 0$.

The root ξ of polynomial (1) is a *root of multiplicity s* (or an *s-multiple root*) if the polynomial itself and all of its derivatives up to the order $s - 1$ vanish at the point $x = \xi$ and the sth derivative does not vanish at that point:

$$P(\xi) = P'(\xi) = \ldots = P^{(s-1)}(\xi), \quad P^{(s)}(\xi) \neq 0.$$

If the root of a polynomial is of multiplicity $s = 1$, then it is a *simple* root.

Example 1. Determine the multiplicity of the root $x = -2$ of the polynomial

$$P(x) = x^4 + 7x^3 + 18x^2 + 20x + 8.$$

△ We seek the first derivative $P'(x) = 4x^3 + 21x^2 + 36x + 20$ and calculate $P'(-2) = 0$. Then we find the second derivative $P''(x) = 12x^2 + 42x + 36$. We have $P''(-2) = 0$. Next we find the third derivative $P'''(x) = 24x + 42$. We have $P'''(-2) = -6 \neq 0$.

Thus the polynomial itself as well as its first and second derivatives vanish at the point $x = -2$ whereas its third derivative is nonzero at that point. This means that the root $\xi = -2$ has multiplicity $s = 3$. ▲

The principal properties of the roots of an algebraic polynomial are presented in 5.8.

Consider the problem of calculating polynomial (1) at a point $x = x^*$. To solve this problem, we represent the polynomial in the form

$$P_n(x) = a_n + x(a_{n-1} + x(a_{n-2} + \ldots + x(a_1 + xa_0)\ldots)). \tag{2}$$

To find the value of $P_n(x^*)$, we calculate, in succession,

$$\begin{aligned} b_0 &= a_0, \\ b_1 &= a_1 + x^* b_0, \\ b_2 &= a_2 + x^* b_1, \\ &\cdots\cdots\cdots \\ b_n &= a_n + x^* b_{n-1}. \end{aligned} \tag{3}$$

We can see from the expressions for b_i that every successive coefficient b_i results from the addition of the corresponding coefficient a_i to the product of the predecessor b_{i-1} by x^*.

When calculating $P_n(x^*)$ and the coefficients b_i, without using a computer it is convenient to use the following table:

Table 4.1

a_0	a_1	a_2	a_3	...	a_{n-1}	a_n
	x^*b_0	x^*b_1	x^*b_2	...	x^*b_{n-2}	x^*b_{n-1}
b_0	b_1	b_2	b_3	...	b_{n-1}	b_n

In the first row we write the coefficients of the polynomial $P_n(x)$ (we write the negative coefficients with the minus sign and omit the plus sign before the positive coefficients). We immediately transfer $b_0 = a_0$ to the third row. Then we multiply every coefficient b_i by x^* and write the result under the next coefficient a_{i+1}. We sum up the numbers in the first and the second row and write the result b_{i+1} in the third row.

By virtue of the construction of the calculation process it is evident that $b_n = P_n(x^*)$.

It is easy to calculate the number of operations needed to find $P_n(x^*)$: with the use of formula (1) we must perform $2n - 1$ multiplications and n additions and with the use of formula (2) we must make m additions and n multiplications.

Thus formula (2) makes it possible to economize and almost half the number of multiplications. And if we take into consideration the fact that the operation of multiplication requires much more time than the operation of addition, we see that the use of calculating scheme (2) is rather efficient.

Scheme (2) is known as **Horner's scheme**.

Example 1. Using Horner's scheme, calculate the value of the polynomial $P_5(x) = x^5 + 3x^4 - 2x^3 + x^2 - x + 1$ for $x = 3$.

△ We construct Horner's scheme for this polynomial:

1	3	−2	1	−1	1
	3	18	48	147	438
1	6	16	49	146	$439 = P_5(3)$

Thus $P_5(3) = 439$. ▲

It turns out that when realizing Horner's scheme we not only calculate $b_n = P_n(x^*)$ but also find the coefficients b_i $(i = 0, 1, \ldots, n-1)$ of the polynomial

$$P^{(1)}_{n-1}(x) = b_0x^{n-1} + b_1x^{n-2} + \ldots + b_{n-1}, \tag{4}$$

which is the quotient of the division of the polynomial $P_n(x)$ by the binomial $x - x^*$:

$$P_n(x) = (x - x^*)\,P^{(1)}_{n-1}(x) + P_n(x^*). \tag{5}$$

Indeed, removing the brackets on the right-hand side of relation (5), collecting terms and taking into account that $P_n(x^*) = b_n$, we obtain

$$P_n(x) = b_0x^n + (b_1 - x^*b_0)\,x^{n-1} + (b_2 - x^*b_1)\,x^{n-2}$$
$$+ \ldots + (b_{n-1} - x^*b_{n-2})\,x + b_n - x^*b_{n-1}.$$

From this we have

$$a_0 = b_0,$$
$$a_1 = b_1 - x^*b_0,$$
$$a_2 = b_2 - x^*b_1,$$
$$a_n = b_n - x^*b_{n-1},$$

which is in complete accord with relations (3) and thus proves representation (5).

We thus arrive at the following theorem.

Bezout's theorem. *The remainder of the division of the polynomial $P_n(x)$ by the binomial $x - x^*$ is equal to the value of the polynomial for $x = x^*$.*

Example 2. Calculate the value of the polynomial $P_4(x) = 12x^4 + 19x^3 - 4$ at the point $x^* = -2$ and determine the coefficients of the polynomial $P_3^{(1)}(x)$ which is the quotient of the division of $P_4(x)$ by the binomial $x + 2$.

△ We construct Horner's scheme for the polynomial:

12	19	0	0	−4
	−24	10	−20	40
12	−5	10	−20	36

Thus $P_4(-2) = 36$ and $P_3^{(1)}(x) = 12x^3 - 5x^2 + 10x - 20$. ▲

Another example of the use of Horner's scheme is a change of variable in a polynomial.

Assume that in polynomial (1) we have to pass to a new variable y which is in a linear relationship with the variable x: $x = y + x^*$, or $y = x - x^*$.

This means that we have to find the coefficients of the polynomial

$$P_n(y + x^*) = A_0 y^n + A_1 y^{n-1} + \ldots + A_{n-1} y + A_n. \quad (6)$$

We can show that

$$A_n = P_n(x^*),$$
$$A_{n-1} = P_{n-1}^{(1)}(x^*),$$
$$A_{n-2} = P_{n-2}^{(2)}(x^*),$$
$$A_0 = P_0^{(n)},$$

where $P_n(x)$ is the given polynomial (1) and the other polynomials

$$P_{n-k}^{(k)}(x) = b_0^{(k)} x^{n-k} + b_1^{(k)} x^{n-k-1} + \ldots + b_{n-k}^{(k)}$$
$$(k = 1, 2, \ldots, n)$$

are defined by the relations

$$P_n(x) = (x - x^*) P_{n-1}^{(1)}(x) + P_n(x^*),$$
$$P_{n-1}^{(1)}(x) = (x - x^*) P_{n-2}^{(2)}(x) + P_{n-1}^{(1)}(x^*),$$
$$P_1^{(n-1)}(x) = (x - x^*) P_0^{(n)} + P_1^{(n-1)}(x^*).$$

Thus we get the following simple algorithm for finding A_j $(j = n, n-1, \ldots, 1, 0)$:

1°. Using Horner's scheme, we calculate the coefficient $A_n = P_n(x^*)$ and also the coefficients $b_i^{(1)}$ $(i = 0, 1, \ldots, n-1)$ of the polynomial $P_{n-1}^{(1)}(x)$.

2°. Applying Horner's scheme to the polynomials $P_{n-1}^{(1)}(x), P_{n-2}^{(2)}(x), \ldots, P_{n-k}^{(k)}(x)$, we calculate the coefficients $A_{n-k} = P_{n-k}^{(k)}(x^*)$ and also the coefficients $b_i^{(k+1)}$ $(i = 0, 1, \ldots, n-k-1)$ of the polynomial $P_{n-k-1}^{(k+1)}(x)$. The process is completed when k becomes equal to $n-1$. Then we set $A_0 = P_0^{(n)} = b_0^{(n)} = a_0$. We have thus found all the coefficients A_j.

This algorithm is often called the **generalized Horner's scheme.**

Example 3. In the polynomial $P_4(x) = 12x^4 + 19x^3 - 4$ pass to a new variable $y = x + 2$.

△ Using relations (5) and (6), we compose a table:

12	19 -24	0 10	0 -20	-4 40
12	-5 -24	10 58	-20 -136	$36 = A_4$
12	-29 -24	68 106	-156	$= A_3$
12	-53 -24	174	$= A_2$	
12	$-77 = A_1$			

$12 = A_0$

Thus the polynomial $Q_4(y) = P_4(y-2)$ has the form

$$Q_4(y) = 12y^4 - 77y^3 + 174y^2 - 156y + 36. \quad ▲$$

4.2. Calculating the Values of Analytic Functions

The calculation of the values of an analytic function is often based on representing it as a quickly converging Taylor's series which is in many cases a convenient instrument for calculating the values of this function at the points belonging to the domain of convergence of the series.

Assume that we have to find the value of the function $f(x)$ analytic on the interval $[a, b]$ at the point $x = x^*$, $x^* \in [a, b]$ with a specified limiting absolute error ε.

Taylor's formula with the remainder in Lagrange's form for the function $f(x)$ in the neighbourhood of the point $x = c$, $c \in [a, b]$ has the form

$$f(x) = f(c) + (x-c)\frac{f'(c)}{1!}$$

$$+ \dots + (x-c)^n \frac{f^n(c)}{n!} + (x-c)^{n+1} \frac{f^{(n+1)}(\bar{c})}{(n+1)!},$$

where $\bar{c} = c + \theta(x-c)$, $0 < \theta < 1$.

Consequently, the problem of computing $f(x^*)$ reduces to the calculation of

$$f(x^*) = S_n(x^*) + R_n(x^*),$$

where $S_n(x^*)$ is the nth partial sum of the series:

$$S_n(x^*) = \sum_{i=0}^{n} (x^*-c)^i \frac{f^{(i)}(c)}{i!}$$

$$(0! = 1,\ f^{(0)}(c) = f(c)),$$

and $R_n(x^*)$ is the value of the remainder $R_n(\cdot)$ for $x = x^*$:

$$R_n(x^*) = (x^*-c)^{n+1} \frac{f^{n+1}(\bar{c}^*)}{(n+1)!},$$

$$\bar{c}^* = c + \theta^*(x^*-c), \quad 0 < \theta^* < 1.$$

The algorithm of the solution of this problem is the following.

1°. On the interval $[a, b]$ we choose a point $x = c$ as close to the point $x = x^*$ as possible and such that the function $f(x)$ itself and its derivatives can be easily calculated for $x = c$.

2°. We represent ε as the sum

$$\varepsilon = \varepsilon_1 + \varepsilon_2 + \varepsilon_3, \tag{1}$$

where ε_1 is the residual error (the error of the method), ε_2 is the limiting absolute error of calculation of $S_n(x^*)$, ε_3 is the limiting absolute error of rounding off of the result. Generally speaking, ε_1, ε_2 and ε_3 can be arbitrary positive numbers satisfying condition (1).

In practial calculations ε is usually given in the form $\varepsilon = 10^{-m}$, where m is an integer. Then we usually assume that $\varepsilon_3 = 0.5\cdot 10^{-m}$ and $\varepsilon_1 = \varepsilon_2 = 0.25\cdot 10^{-m}$. If there is no error of the final rounding-off, then we assume that $\varepsilon_1 = \varepsilon_2 = 0.5\cdot 10^{-m}$, $\varepsilon_3 = 0$.

3°. We choose the number of terms in S_n such that the inequality

$$|f(x^*) - S_n(x^*)| = R_n(x^*) \leqslant \varepsilon_1$$

is satisfied.

4°. Each term in S_n is calculated so that the approximate value of $\bar{S}_n$ differs from the exact value of S_n by not more than ε_2. For that purpose, each term of $\bar{S}_n$ is usually calculated with an absolute error $\varepsilon_2/(n+1)$.

5°. The approximate sum $\bar{S}_n$ obtained in 4° is rounded off (if $\varepsilon_3 \neq 0$) to the value $\bar{\bar{S}}_n$.

6°. The solution of the problem is written in the form $f(x^*) = \bar{\bar{S}}_n \pm \varepsilon$.

Example. Calculate $e^{2.25}$ with an accuracy to within $\varepsilon = 0.01$.

△ Taylor's formula with the remainder in Lagrange's form for the function e^x in the neighbourhood of the point $x = 0$ has the form

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \ldots + \frac{x^n}{n!} + e^{\theta x}\frac{x^{n+1}}{(n+1)!}, \quad 0 < \theta < 1.$$

Since for large x Taylor's series of the function e^x converges slower, it is expedient to calculate the value of $e^{2.25}$ as the product $e^2 \cdot e^{0.25}$

The quantity e^2 can be easily calculated with any degree of accuracy and we can assume that practically the error of its calculation is zero.

Let us assume that the errors of rounding off and calculation of $e^2 \cdot e^{0.25}$ are equal to 0.005. Then the error of calculation of $e^{0.25}$ is $\bar{\varepsilon} = 0.005/e^2 \cong 0.0006$.

Thus we have reduced the calculation of $e^{2.25}$ with an accuracy of $\varepsilon = 0.01$ to the calculation of

$$e^{0.25} = 1 + 0.25^2 + \frac{0.25^2}{2!} + \ldots + \frac{0.25^n}{n!} + e^{0.25\theta}\frac{0.25^{n+1}}{(n+1)!}$$

with an accuracy of $\bar{\varepsilon} = 0.0006$.

Assume, furthermore, that $\varepsilon_1 = \varepsilon_2 = 0.0003$. We shall find n using the remainder of Taylor's formula for e^x when $x = 0.25$:

$$e^{0.25\theta}\frac{0.25^{n+1}}{(n+1)!} \leqslant 0.0003;\quad 0<\theta<1.$$

Taking into account that $e^{0.25\theta} < e^{0.25} < 1.5$, we get $n \geqslant 3$. Consequently, we must calculate the sum

$$S_3 = 1+0.25+\frac{0.25^2}{2!}+\frac{0.25^3}{3!}$$

with an absolute error 0.0003. Since we have calculated the first two terms with an absolute accuracy, it is sufficient to calculate the second and the third term with the limiting absolute error of 0.0001 each. Performing the necessary calculations, we find that

$$\overline{S}_3 = 1 + 0.25 + 0.0312 + 0.0026 = 1.2838.$$

Multiplying the result by e^2, we get $e^2\overline{S}_3 = 9.4860\ldots$. Finally, rounding off to the hundreds place, we get $e^{2.25} = 9.49 \pm 0.01$. ▲

We can similarly calculate the values of trigonometric functions (sine and cosine). The corresponding Taylor's formulas with the remainder in Lagrange's form are

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!}$$

$$- \ldots + (-1)^{n-1}\frac{x^{2n-1}}{(2n-1)!} + (-1)^n\frac{x^{2n+1}}{(2n+1)!}\cdot\theta,\ 0<\theta<1, \tag{2}$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!}$$

$$- \ldots + (-1)^{n-1}\frac{x^{2n-2}}{(2n-2)!} + (-1)^n\frac{x^{2n}}{(2n)!}\cdot\theta,\ 0<\theta<1. \tag{3}$$

The argument in formulas (2) and (3) must be in radians and belong to the interval $[0, \pi/4]$. In that case the convergence of the corresponding series will be sufficiently fast. The reduction formulas and the known relation of trigonometric functions will make the argument belong to the indicated interval.

Thus, for instance, to calculate the value of sin 2.53, we must use the reduction formula $\sin 2.53 = \sin(\pi - 2.53)$; the argument $\pi - 2.53$ belongs to the interval $[0, \pi/4]$. To calculate the value of cos 1.27, we must use the formula $\cos 1.27 = \sin(\pi/2 - 1.27)$ and compute the value of the sine since $(\pi/2 - 1.27) \in [0, \pi/4]$.

We write the condition for the choice of the number of terms n in calculating the value of the sine as

$$\frac{x^{2n+1}}{(2n+1)!} \leqslant \varepsilon_1, \quad x \in [0, \pi/4], \tag{4}$$

and in the calculation of the cosine as

$$\frac{x^{2n}}{(2n)!} \leqslant \varepsilon_1, \quad x \in [0, \pi/4]. \tag{5}$$

The formula for the expansion into series

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots + (-1)^{n-1}\frac{x^n}{n} + \ldots; \quad -1 < x \leqslant 1 \tag{6}$$

is hardly convenient for calculating the values of a logarithmic function because of the slow convergence for $|x|$ close to unity. In addition, this formula does not allow us to calculate the logarithms of numbers larger than two.

To accelerate the convergence and expand the domain of applicability, we transform formula (6) as follows. Replacing x by $-x$ on its right-hand and left-hand sides, we have

$$\ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \ldots - \frac{x^n}{n} - \ldots$$

Subtracting the initial relation from the resulting one we find that

$$\ln\frac{1-x}{1+x} = -2\left(x + \frac{x^3}{3} + \frac{x^5}{5} + \ldots + \frac{x^{2n-1}}{2n-1} + \ldots\right).$$

Setting now $(1-x)/(1+x) = z$ and bearing in mind that $x = (1-z)/(1+z)$, we obtain the initial formula for calculating the natural logarithm of any number z belonging to the interval $(0, \infty)$:

$$\ln z = -2\left[\frac{1-z}{1+z} + \frac{1}{3}\left(\frac{1-z}{1+z}\right)^3 + \frac{1}{5}\left(\frac{1-z}{1+z}\right)^5 + \ldots\right]. \tag{7}$$

In practical computations, it is convenient to represent the positive number x, whose natural logarithm must be found, in the form

$$x = 2^m \cdot z, \tag{8}$$

where m is an integer and

$$0.5 \leqslant z < 1. \tag{9}$$

Then

$$\ln x = m \ln 2 + \ln z.$$

It is easy to calculate the first term for the known m and $\ln 2 = 0.69314718\ldots$, and the second term can be found from formula (7). By virtue of inequality (9), the quantity

$$\xi = (1 - z)/(1 + z) \tag{10}$$

varies in the limits

$$0 < \xi \leqslant 1/3, \tag{11}$$

and this ensures fast convergence of series (7).

The final result is

$$\ln x = m \ln 2 - 2\left(\xi + \frac{\xi^3}{3} + \frac{\xi^5}{5} + \ldots + \frac{\xi^{2n-1}}{2n-1}\right) - R_n. \tag{12}$$

We estimate the remainder:

$$R_n = 2\left(\frac{\xi^{2n+1}}{2n+1} + \frac{\xi^{2n+3}}{2n+3} + \ldots\right) < \frac{2}{2n+1} \cdot \frac{\xi^{2n+1}}{1-\xi^2}.$$

Using inequality (11), we can reduce this estimate to the form convenient for comparison with the corresponding terms of series (7):

$$R_n < \frac{9}{4}\,\frac{\xi^{2n+1}}{2n+1} \leqslant \frac{\xi^{2n-1}}{4(2n+1)}. \tag{13}$$

To compute the value of a decimal logarithm, we must use the formula

$$\log x = M \ln x,$$

where $M = \log e = 0.434294481903252\ldots$.

4.3. The Iterative Method of Calculating the Value of a Function

Consider one more artificial technique of calculating the value of the function

$$y = f(x), \tag{1}$$

continuous on the interval $[a, b]$, at the point $x = x^*$, $x^* \in [a, b]$.

This technique is based on the use of Newton's method for solving algebraic and transcendental equations and consists in the following.

1°. Function (1) is written in implicit form and the value of x^* is substituted for x in the resulting expression:

$$F(x^*, y) = 0. \tag{2}$$

The required value of the function $y^* = f(x^*)$ is precisely the solution of this equation.

2°. Equation (2) is solved by Newton's method, for which purpose the initial approximation y_0 is chosen such that the condition

$$F(x^*, y_0) F''_{yy}(x^*, y_0) > 0 \tag{3}$$

is satisfied and each successive approximation y_n $(n = 1, 2, 3, \ldots)$ is calculated from the formula

$$y_n = y_{n-1} - \frac{F(x^*, y_{n-1})}{F'_y(x^*, y_{n-1})} \quad (n = 1, 2, 3, \ldots). \tag{4}$$

The convergence of the iterative process (4) is ensured when the function $F(x^*, y)$ satisfies the conditions of convergence of Newton's method.

Note that there are numerous ways to represent function (1) in implicit form (2), from which we must choose the representation such that the iterative process (4) would be convergent and the rate of convergence would be sufficiently high.

An important example of this technique is the calculation of the function $y = \sqrt[m]{x}$ $(m = 2, 3, \ldots)$ in the interval $(0, \infty)$.

It is expedient to use the expression $y^m - x$ as $F(x^*, y)$ for this function. Then condition (3) is reduced to the form

$$y_0 > \sqrt[m]{x^*} \tag{3'}$$

and the iterative process (4) to the form

$$y_n = y_{n-1} - \frac{y_{n-1}^m - x^*}{m y_{n-1}^{m-1}}. \tag{4'}$$

Note that by virtue of the properties of the function $y^m - x^*$, the iterative process (4′) converges not only when condition (3′) is fulfilled but also for any positive initial approximation $y_0 > 0$. In this case the condition $y_n > \sqrt[m]{x^*}$ $(n = 1, 2, \ldots)$ will be fulfilled for all the successive approximations.

We can estimate the error of the approximation y_n as follows:

$$\Delta y = y_n - \sqrt[m]{x^*} < \sqrt{\frac{y_{n-1}^m}{x^*}}\,(y_{n-1} - y_n),$$

or

$$\Delta y_n = y_n - \sqrt[m]{x^*} < \frac{m-1}{2} \cdot \frac{y_{n-1}^{m-1}}{x^*}(y_{n-1} - y_n)^2. \tag{5}$$

Since the case of $m = 2$ is encountered much more frequently, we shall give for it a more accurate estimate of the error.

We represent x^* in the form $x^* = 2^k \bar{x}$, where k is an integer $\bar{x}^* \in [0.5, 1)$. Then, setting $y_0 = 2^{E(k}$, we obtain

$$\Delta y_n = y_n - \sqrt{x^*} < \frac{25}{12} y_1 \left(\frac{1}{5}\right)^{2^n} \leqslant \frac{25}{8} y_0 \left(\frac{1}{5}\right)^{2^n}.$$

Example. Calculate $\sqrt[3]{34}$ with an accuracy of $\varepsilon = 10^{-4}$.

△ In accordance with inequality (3′) we choose $y_0 = 3.4 > \sqrt[3]{34}$. Then, using formula (4′) for $x^* = 34$ and $y_0 = 3.4$, we find, in succession, y_n and calculate Δy_n from formula (5):

$$y_1 = 3.4 - \frac{3.4^3 - 34}{3 \cdot 3.4^2} = 3.247, \quad y_1 = 0.01;$$

$$y_2 = 3.247 - \frac{3.247^3 - 34}{3 \cdot 3.247^2} = 3.23964, \quad \Delta y_2 = 2 \cdot 10^{-5}.$$

Thus $\sqrt[3]{34} = 3.23964 \pm 0.00002$. ▲

Exercises

1. Using Horner's scheme, perform the division of the polynomial $P_5(x) = x^5 + 3x^4 - 2x^3 - x + 1$ by the binomial $x - 3$.

2. Find out whether $\xi = 1$ is a root of the equation $x^3 + 2x^2 - 3 = 0$.

3. Use the expansion in a power series with an accuracy of $\varepsilon = 0.001$ to calculate (a) sin 25°, (b) cos 20°, (c) ln 4, (d) $\sqrt[5]{e}$, (e) cos 36°, (f) sin 18°.

4. Use the iterative method to calculate, with an accuracy of $\varepsilon = 0.0005$, (a) $\sqrt{12}$, (b) $\sqrt{56}$, (c) $\sqrt{42}$.

Chapter 5

Methods of Solving Nonlinear Equations

5.1. Algebraic and Transcendental Equations

In practical problems we often have to solve equations. We can represent every equation in one unknown in the form

$$\varphi(x) = g(x), \tag{1}$$

where $\varphi(x)$ and $g(x)$ are given functions defined on a number set X called the *domain of permissible values of the equation.*

We can write an equation in one unknown as

$$f(x) = 0. \tag{2}$$

Indeed, transferring $g(x)$ to the left-hand side of equation (1), we get an equation $\varphi(x) - g(x) = 0$ which is equivalent to (1). If we designate the left-hand side of the last equation as $f(x)$, we get equation (2).

The set of values of the variable x for which equation (1) turns into an identity is a *solution* of this equation and every value of x from that set is a *root* of the equation.

For instance, the equation $x^2 = 2 - x$ has roots $x_1 = -2$ and $x_2 = 1$. Substituting -2 and 1 into the given equation for x, we get identities $(-2)^2 = 2 - (-2)$, i.e. $4 \equiv 4$; $1^2 = 2 - 1$, i.e. $1 \equiv 1$.

To solve an equation is to find the set of all roots of the equation. It may be finite or infinite. Thus the equation considered above has two roots. The equation $\sin x = 0$ has a solution $x = \pi n$ ($n = 0, \pm 1, \pm 2, \ldots$). Assigning different values to n, we get an infinity of roots.

A set of several equations in several unknowns is a *system of equations* (the unknown denoted by the same letter in all the equations must mean the same unknown quantity).

A *solution of a system of equations* in several unknowns is the set of values of the unknowns which turns every equation of the system into an identity.

For example, the system

$$\begin{cases} x^2+y=5,\\ x+y^2=3 \end{cases}$$

has a solution $x=2$, $y=1$ since for these values of the unknowns the equations of the system turn into identities: $4+1\equiv 5$, $2+1\equiv 3$.

To solve a system of equations is to find the set of all its solutions or to show that it has no solutions.

According as what functions enter into equations (1) and (2) equations are divided into two large classes, algebraic and transcendental equations.

A function is *algebraic* if, in order to get the values of the function proceeding from the given value of x, we must perform arithmetic operations and raise to a power with a rational exponent. (The operation of extracting a root can be represented as the operation of raising to a power with the exponent $1/n$.)

An algebraic function is *rational* with respect to the variable x if no operations except for addition, subtraction, multiplication, division and raising to an integral power are performed which involve x.

For example,

$$f_1(x)=x^3+15x^2-1200x+4,\quad f_2(x)=\frac{2}{x-8}+\frac{45}{x+5},$$

$$f_3(x)=(x-4)(x+5),\qquad f_4(x)=\frac{3}{x+7}+\frac{4x+3}{3x^2+5}.$$

If the variable x does not enter into a rational function as a divisor or does not enter into the expression which is a divisor, then the rational function is an *integral*, or *entire*, *rational function*.

For instance, the functions

$y=a_0x^n+a_1x^{n-1}+\ldots+a_n$ (n is a natural number or zero, $a_0, a_1, \ldots, a_n$ are any real numbers with $a_0\neq 0$),

$$f(x)=\frac{x^2}{4}+\frac{x+8}{3}$$

are entire rational functions. An entire rational function is defined throughout the number axis.

If, in a rational function, a division by the variable x is encountered at least once or the variable x enters into the expression which is a divisor, then the function is a *rational fractional function.*

For instance, the function

$$y=\frac{b_0x^m+b_1x^{m-1}+\ldots+b_m}{a_0x^n+a_1x^{n-1}+\ldots+a_n},$$

where m is a natural number or zero, n is a natural number, $a_0, a_1, \ldots, b_0, b_1, \ldots$ are any real numbers ($a_0 \neq 0$, $b_0 \neq 0$), is a rational fractional function.

A rational fractional function is defined on the entire number axis except for the points at which the denominator vanishes.

A function is *irrational* if, to obtain the value of the function proceeding from the given value of x, we must perform, in addition to the four arithmetic operations (all or some of them), the operation of extracting a root. In that case, the function is irrational only if the argument x is under the sign of the radical.

Thus the function

$$y=\frac{3x^2-4x+\sqrt[3]{x-1}}{7x-4}$$

is irrational whereas the function

$$y=\sqrt{\frac{1+\sqrt{3}}{4}}\,x^2+\frac{\sqrt{5}}{2}\,x+4$$

is not irrational since x is not under the sign of the radical.

We have mentioned earlier that all rational and irrational functions belong to the class of algebraic functions.

Transcendental functions is another large class of functions. They include all nonalgebraic functions, i.e. an exponential function a^x, a logarithmic function $\log_a x$, trigonometric functions $\sin x$, $\cos x$, $\tan x$, $\cot x$, inverse trigonometric functions $\arcsin x$, $\arccos x$, $\arctan x$, $\operatorname{arccot} x$ and others.

If the notation of an equation includes only algebraic functions, then the equation is *algebraic*.

For instance, the equations

$$x^5 - 4 = 0, \quad x^4 - 3x^3 + 5x^2 - x + 1 = 0$$

are algebraic.

An algebraic equation can be reduced to the form

$$a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \ldots + a_{n-1}x + a_n = 0. \qquad (3)$$

Therefore, when we say "an algebraic equation", we usually have in mind an equation of form (3).

If equation (3) has been obtained by means of the transformation of an equation which contained a rational fractional or irrational function, then we must take into account the fact that these functions are defined not throughout the number axis.

For instance, the equation

$$\sqrt{x-2} + \sqrt{6-x} = 3$$

assumes the form

$$4x^2 - 16x - 47 = 0$$

after the function is rationalized. However, the initial equation is defined not on the entire number axis but only for x which belong to the interval [2, 6].

The numbers a_0, a_1, . . ., a_n are the *coefficients* of equation (3)' and may be real as well as complex. In what follows, we consider algebraic equations of form (3) only with real coefficients.

To solve an equation in one unknown is to find its roots, i.e. the values of x which turn the equation into an identity. The roots of an equation may be real or nonreal (complex).

We can find the exact roots of an equation only in exceptional cases, usually when there is a simple formula for the calculation of the values of the roots which makes it possible to express them in terms of the known quantities.

Thus, to find the roots of a quadratic equation of the form $x^2 + px + q = 0$, we use the formula

$$x_{1,2} = -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q}. \qquad (4)$$

To solve a cubic equation of the form $x^3 + px + q = 0$, we use the formula

$$x = \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}}. \qquad (5)$$

However, it is difficult to use this formula in practice since it requires the use of complex numbers.

There is also a formula for solving a fourth-degree equation but it is so complicated that it is not employed in practice and we shall not consider it here.

Abel, a Norwegian mathematician, has proved that for $n \geqslant 5$ there is no formula which would express the solution of the algebraic equation (3) by means of the arithmetic operations and the extraction of roots. Only in some special cases there may be formulas for solving algebraic equations whose degree is higher than four.

In addition, the coefficients of some equations are approximate numbers and, consequently, we cannot pose the problem of finding exact roots.

Therefore, of considerable importance are the methods of approximation of the roots of the equation $f(x) = 0$.

In many practical problems it is not always necessary to find the exact solution of an equation. The problem of finding the roots is considered to be solved when the roots are calculated with the specified degree of accuracy.

Then how must we understand the statement "the root has been calculated with the specified degree of accuracy"? Let ξ be a root of an equation and $\bar{x}$ its approximate value with an accuracy of ε. This means that $|\xi - \bar{x}| \leqslant \varepsilon$. If it is established that the required root ξ is between the numbers a and b, i.e. $a < \xi < b$, with $b - a \leqslant \varepsilon$, then the numbers a and b are approximate values of the root ξ, by excess and by defect respectively, with an accuracy of ε since $|\xi - a| < b - a \leqslant \varepsilon$ and $|\xi - b| < b - a \leqslant \varepsilon$. We can assume any number included between a and b to be an approximate value of the root ξ with an accuracy of ε.

For instance, if the root ξ is between 3.228 and 3.229 (i.e. $3.228 < \xi < 3.229$), then, with an accuracy of 0.001, we can take the number 3.228, 3.229 and any number included between them to be an approximate value of the root.

In this chapter we consider the methods of approximation of equations and systems of equations. Some of them can be employed in the solution of both transcendental and algebraic equations. Other methods can be used to solve only algebraic equations.

5.2. Separation of Roots

The process of seeking approximate values of the roots of an equation can be divided into two stages: (1) separation of the roots, and (2) computation of the values of the roots with the specified degree of accuracy.

This section is devoted to the first stage, i.e. separation of roots.

The root ξ of the equation $f(x) = 0$ is considered to be *separated* on the interval $[a, b]$ if the equation $f(x) = 0$ has no other roots on this interval.

To separate roots is to divide the whole domain of permissible values into intervals in each of which there is one root. There are two methods of separating roots, graphical and analytical.

Graphical method of separating roots. *1st technique.* It is easy to separate roots if the graph of the function $y = f(x)$ is constructed. The points of intersection of the graph and the x-axis yield the values of the root and it is easy, using the graph, to find two numbers a and b which include only one root between them (Fig. 5.1).

2nd technique. All terms of an equation are divided into two groups, one of them is written on the left-hand side of the equation and the other on the right-hand side, i.e. the equation is represented as $\varphi(x) = g(x)$. Then the graphs of two functions, $y = \varphi(x)$ and $y = g(x)$, are constructed. The abscissas of the points of intersection of the graphs of these two functions are the roots of the equation. Let the point of intersection of the graphs have abscissa x_0 and the ordinates of the two graphs at that points be equal, i.e. $\varphi(x_0) = g(x_0)$. It follows from

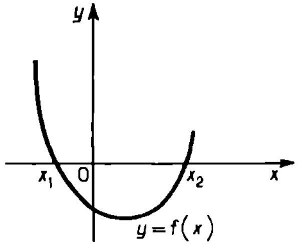


Fig. 5.1

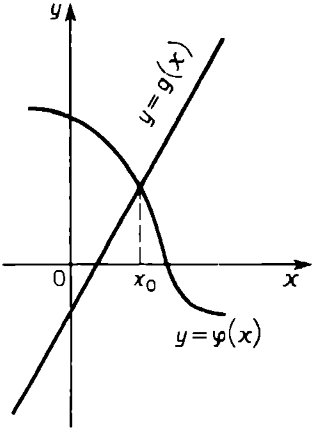


Fig. 5.2

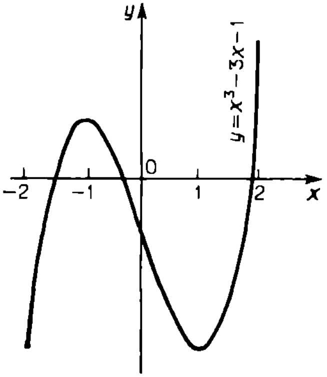


Fig. 5.3

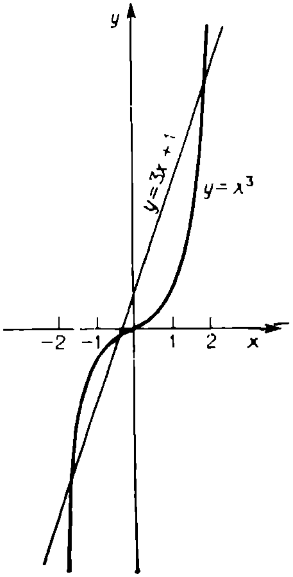


Fig. 5.4

this equality that x_0 is a root of the equation (Fig. 5.2). The numbers a and b which include the root between them can be determined from the graph.

Example 1. Use graphical means to find the integers between which the roots of the equation $x^3 - 3x - 1 = 0$ are included.

△ *1st technique.* We construct the graph of the function $y = x^3 - 3x - 1$ (Fig. 5.3) and find the abscissas of the points of intersection of the graph and the x-axis. The curve cuts the x-axis at three points, and, consequently, the equation has three real roots (note that a third-degree algebraic equation has either one or three real roots). We can see from the drawing that the roots belong to the intervals $[-2, -1]$, $[-1, 0]$ and $[1, 2]$.

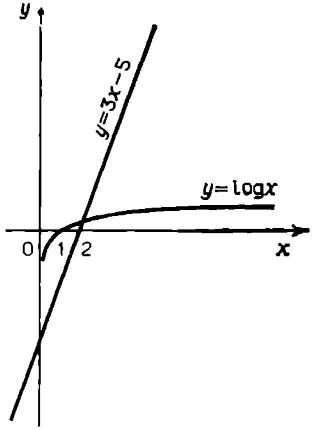


Fig. 5.5

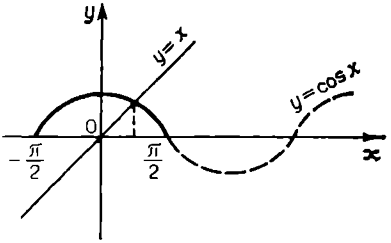


Fig. 5.6

2nd technique. We represent the equation in the form $x^3 = 3x + 1$ and construct the graphs of the functions $x = x^3$ and $y = 3x + 1$ (Fig. 5.4). The abscissas of the points of intersection of the graphs of these functions are roots of the equation. The intervals of the separation of the roots can be easily found from the drawings. ▲

Example 2. Use graphical means to separate the roots of the equation $\log x - 3x + 5 = 0$.

△ We rewrite the equation as follows: $\log x = 3x - 5$. The functions on the left-hand and right-hand sides of the equation have a common domain of definition, the interval $0 < x < +\infty$. We shall therefore seek the roots in this interval.

We construct the graphs of the functions $y = \log x$ and $y = 3x - 5$ (Fig. 5.5). The straight line $y = 3x - 5$ cuts the logarithmic curve at two points. It is difficult to show on the drawing the intersection of the graphs of these functions at the first point. However, taking into consideration that the lower branch of the logarithmic curve approaches the y-axis indefinitely, we can assume that the intersection of the graphs will be close to the point of intersection of the graph of the function $y = 3x - 5$ and the y-axis. Thus the roots lie on the intervals $[0, 0.5]$ and $[1, 2]$. ▲

Example 3. Use graphical means to separate the roots of the equation $x - \cos x = 0$.

△ We rewrite the equation in the form $x = \cos x$ and construct the graphs of the functions $y = x$ and $y = \cos x$ in the interval $-\pi/2 \leqslant x \leqslant \pi/2$. The graphs of the functions intersect at one point; in this interval the equation has one root. Taking into account the properties of the functions $y = x$ and $y = \cos x$, we can make sure that outside of this interval the equation has no roots. If we construct a more precise drawing, we can find that the root is on the interval $[0.6, 0.8]$ (Fig. 5.6). ▲

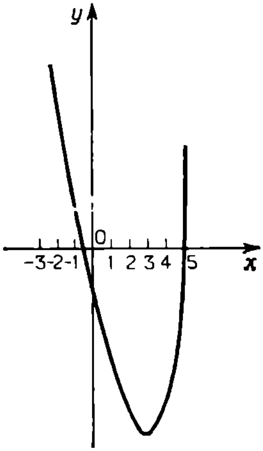


Fig. 5.7

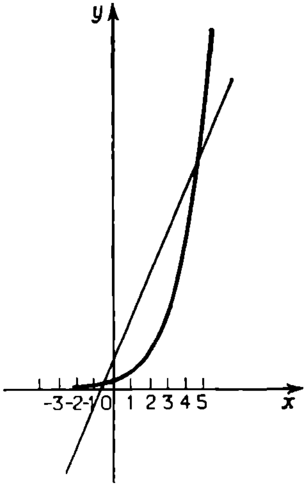


Fig. 5.8

Example 4. Use graphical means to separate the roots of the equation $2^x - 5x - 3 = 0$ employing two techniques.

△ *1st technique.* We construct the graph of the function $y = 2^x - 5x - 3$ and determine the abscissas of the points of intersection of the graph and the x-axis. The curve cuts the x-axis at two points, and, consequently, the equation has two real roots. It can be seen from the drawing that the roots lie on the intervals $[-1, 0]$ and $[4, 5]$ (Fig. 5.7).

2nd technique. We represent the equation in the form $2^x = 5x + 3$ and construct the graphs of the functions $y = 2^x$ and $y = 5x + 3$. We find the abscissas of the points of intersection of the of these functions which are the roots of the given equation. We get the same intervals of separation of the roots $[-1, 0]$ and $[4, 5]$ (Fig. 5.8). ▲

Remark. Assume that the graph of the function $y = f(x)$ has the form given in Fig. 5.9. The curve cuts the abscissa axis three times and, consequently, the equation has three simple roots.

Now if the curve touches the abscissa axis (Fig. 5.10), then the equation has a root of multiplicity two. For instance, the equation $x^3 - 3x + 2 = 0$ has three roots, $x_1 = -2$, $x_2 = x_3 = 1$ (Fig. 5.11).

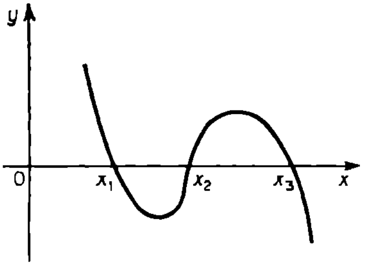


Fig. 5.9

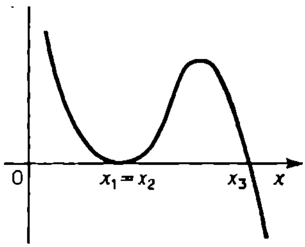


Fig. 5.10

If an equation has a real root of multiplicity three, then, at the point where the curve $y = f(x)$ touches the axis, it has a point of inflection (Fig. 5.12). For instance, the equation $x^3 - 3x^2 + 3x - 1 = 0$ has a root of multiplicity three equal to unity (Fig. 5.13).

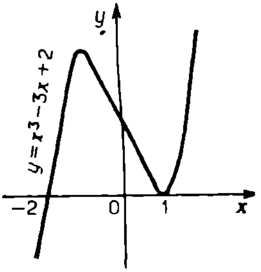


Fig. 5.11

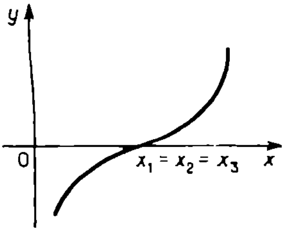


Fig. 5.12

The graphical method of separation of roots is not very precise. It makes it possible to roughly determine the intervals of separation of the roots. Then one of the techniques given below is used to compute the values of the roots.

Analytical method of separating roots. We can separate the roots of the equation $f(x) = 0$ analytically if we use some properties of functions studied in the course of mathematical analysis.

We shall formulate, without proof, some theorems which must be known in order to separate roots.

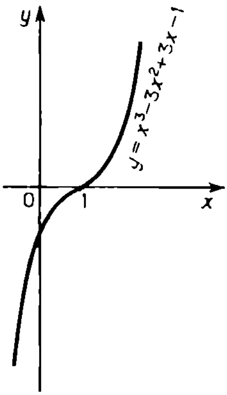


Fig. 5.13

Theorem 1. *If the function* $f(x)$ *is continuous on the interval* $[a, b]$ *and assumes values of unlike signs at the endpoints of this interval, then at least one root of the equation* $f(x) = 0$ *lies within this interval.*

Theorem 2. *If the function* $f(x)$ *is continuous and monotonic on the interval* $[a, b]$ *and assumes values of unlike signs at the endpoints of this interval, then there is a root of the equation* $f(x) = 0$ *within the interval* $[a, b]$ *and that root is unique.*

Theorem 3. *If the function* $f(x)$ *is continuous on the interval* $[a, b]$ *and assumes values of unlike signs at the endpoints of this interval and the derivative* $f'(x)$ *retains sign within the interval, then there is a root of the equation* $f(x) = 0$ *within the interval and that root is unique.*

Here are some data from mathematical analysis which will be needed later on.

If the function $f(x)$ is defined analytically, then the *domain of existence* (*domain of definition*) *of the function* is the set of all the real values of the argument for which the analytical expression defining the function does not loose the numerical sense and assumes only real values.

The function $y = f(x)$ is *monotonic* in a given interval if it satisfies the condition $f(x_2) \geqslant f(x_1)$ or the condition $f(x_2) \leqslant f(x_1)$ for any $x_2 > x_1$ belonging to this interval.

If the continuous function $y = f(x)$ has a derivative at all interior points of the given interval, then the necessary and sufficient condition for the monotonicity of the function in this interval is the satisfaction of the inequality $f'(x) \geqslant 0$ or $f'(x) \leqslant 0$.

Let the function $f(x)$ be continuous on the interval $[a, b]$ and assume values of unlike signs at the endpoints of that interval and let the derivative $f'(x)$ retain sign in the interval (a, b). Then, *if at all points of the interval* (a, b) *the first derivative is positive, i.e.* $f'(x) > 0$, *then the function* $f(x)$ *increases in this interval* (Fig. 5.14 *a*, *c*).

Now if the first derivative is negative at all points of the interval (a, b), *i.e.* $f'(x) < 0$, *then the function decreases in this interval* (Fig. 5.14 *b*, *d*). The root of the function is the abscissa of the point of intersection of the graph of the function $f(x)$ and the x-axis.

Assume that on the interval $[a, b]$ the function $f(x)$ has a second-order derivative which retains sign throughout the interval. Then, *if* $f''(x) > 0$, *then the graph of the function is convex downwards* (Fig. 5.14 *a*, *d*); *now if* $f''(x) < 0$, *then the graph of the function is convex upwards* (Fig. 5.14 *b*, *c*).

The points at which the first derivative of the function is zero and also the points at which it does not exist (say, vanishes), but the function retains its continuity are *critical points* (this test is the necessary condition for extremum).

If the function $f(x)$ is continuous on the interval $[a, b]$ then there are always points on this interval at which it assumes its greatest and least values. The function attains these values either at the critical points or at the endpoints of the interval. Consequently, to find the greatest and the least value of the function on a closed interval, we must (1) find the critical points of the function, (2) calculate the values of the function at the critical points and at the endpoints of the interval $[a, b]$, (3) the greatest value found in item 2 will be the greatest and the least value will be the least value of the function on the

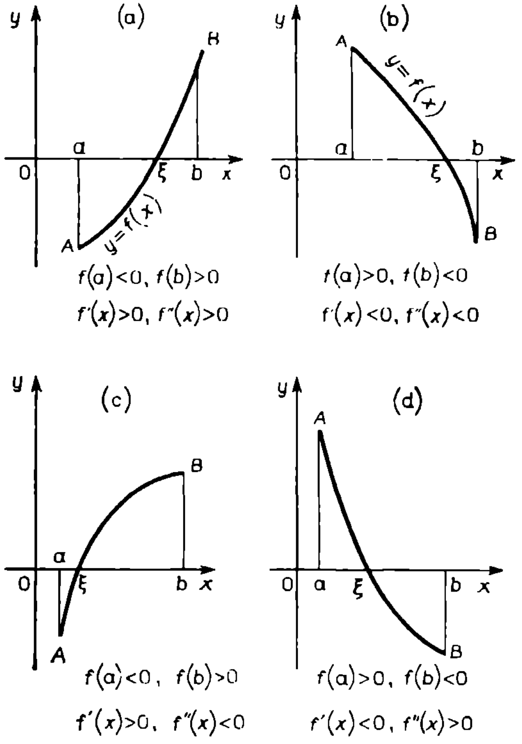


Fig. 5.14

closed interval. In accordance with the aforesaid, we can recommend the following sequence of operations to separate the roots using the analytical method.

(1) find the first derivative $f'(x)$,

(2) compile a table of signs of the function $f(x)$ setting x equal to: (a) the critical values (roots) of the derivative or the values close to them, (b) the boundary values (proceeding from the domain of permissible values of the unknown),

(3) determine the intervals at the endpoints of which the function assumes values of opposite signs. These intervals contain one and only one root each in its interior.

Example 5. Separate the roots of the equation $2^x - 5x - 3 = 0$ using the analytical method.

△ We designate $f(x) = 2^x - 5x - 3$. The domain of definition of the function $f(x)$ is the entire number axis. We seek the first derivative $f'(x) = 2^x \ln 2 - 5$.

We equate this derivative to zero and calculate the root:

$$2^x \ln 2 - 5 = 0, \quad 2^x \ln 2 = 5, \quad 2^x = 5/\ln 2,$$

$$x \log 2 = \log 5 - \log \ln 2,$$

$$x = \frac{\log 5 - \log \ln 2}{\log 2} = \frac{0.6990 + 0.1592}{0.3010} = \frac{0.8582}{0.3010} = 2.85.$$

We compile a table of signs of the function $f(x)$ setting x equal to: (a) the critical values (the roots of the derivative) or the values close to them, (b) boundary values (proceeding from the domain of permissible values of the unknown):

x	$-\infty$	2	3	$+\infty$
Sign of $f(x)$	+	−	−	+

The equation has two roots since the function twice changes sign. We compile a new table with smaller intervals of the separation of the root:

x	−1	0	1	2	3	4	5
Sign of $f(x)$	+	−	−	−	−	−	+

The roots of the equation are in the intervals (−1, 0) and (4, 5). ▲

5.3. Computing Roots with a Specified Accuracy. Trial and Error Method

The preceding section was devoted to the separation of roots, the first stage of solution of algebraic and transcendental equations.

The second stage is the computation of roots with the specified degree of accuracy.

We consider here some ways of refining roots used to solve algebraic and transcendental equations. There are techniques, however, which can only be employed to solve algebraic equations. We shall consider them later on.

Consider an equation $f(x) = 0$, where $f(x)$ is a continuous function. We have to find the root ξ of the equation with an accuracy of ε, where ε is some positive sufficiently small number.

We assume that the root ξ has been separated and lies on the interval $[a, b]$, i.e. there holds an inequality $a \leqslant \xi \leqslant b$. The numbers a and b are approximate values of the root ξ by defect and by excess respectively. The error of these approximations does not exceed the length of the interval $b - a$. If $b - a \leqslant \varepsilon$, then the necessary accuracy of calculations is attained and we can take either a or b as the approximate value of ξ. Now if $b - a > \varepsilon$, then the required accuracy of calculations is not attained and the interval containing ξ must be made narrower, i.e. the numbers $\bar{a}$ and $\bar{b}$ must be chosen so that the inequalities $\bar{a} < \xi < \bar{b}$ and $\bar{b} - \bar{a} < b - a$ are satisfied. The calculations must be terminated when $\bar{b} - \bar{a} \leqslant \varepsilon$ and either $\bar{a}$ or $\bar{b}$ must be taken as the approximate value of the root with an accuracy of ε. It should be pointed out that the value of the root will be more exact when we take the midpoint of the interval, i.e. $c = (\bar{a} + \bar{b})/2$, rather than the endpoints $\bar{a}$ and $\bar{b}$, as the approximate value of the root. In this case the error does not exceed the value $(\bar{b} - \bar{a})/2$.

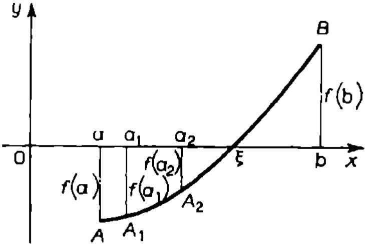


Fig. 5.15

Assume that the root ξ of the equation $f(x) = 0$ [$f(x)$ is a continuous function] has been separated on the interval $[a, b]$, i.e. $f(a) \cdot f(b) < 0$, with $b - a > \varepsilon$. We have to find the value of the root ξ with an accuracy of ε (Fig. 5.15)

On the interval $[a, b]$ we arbitrarily choose a point a_1 which bisects it into two intervals $[a, a_1]$ and $[a_1, b]$. Out of these two intervals we must choose the interval at whose endpoints the function assumes values opposite in sign. In the case being considered $f(a) \cdot f(a_1) > 0$, $f(a_1) \cdot f(b) < 0$, and we must therefore choose the interval $[a_1, b]$. Then, on this narrowed interval, we again arbitrarily choose a point a_2 and find the signs of the products

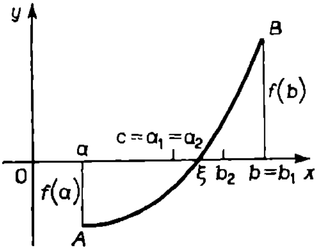


Fig. 5.16

$f(a_1) \cdot f(a_2)$ and $f(a_2) \cdot f(b)$. Since $f(a_2) \cdot f(b) < 0$, we choose the interval $[a_2, b]$. We continue this process until the length of the interval on which the root lies becomes smaller than ε. We obtain the root ξ as the arithmetic mean of the endpoints of the interval obtained; the error of the root does not exceed $\varepsilon/2$.

The method we have considered is known as the trial and error method.

In the form considered above this method is not applicable to calculations on computers. To compose a program and to make calculations on computers, it is used in the form of the so-called **bisection method**.

Assume that the root ξ of the equation $f(x) = 0$ has been separated and lies on the interval $[a, b]$, i.e. $f(a) \cdot f(b) < 0$, with $b - a > \varepsilon$ [here $f(x)$ is a continuous function]. As before, we take an intermediate point on the interval $[a, b]$, not in an arbitrary way, however, but so that it is the midpoint of the interval $[a, b]$, i.e. $c = (a + b)/2$. Then the point c divides the interval $[a, b]$ into two equal intervals $[a, c]$ and $[c, b]$ which are equal in length to $(b - a)/2$ (Fig. 5.16). If $f(c) = 0$, then c

is the exact root of the equation $f(x) = 0$. Now if $f(c) \neq 0$, then out of the two intervals $[a, c]$ and $[c, b]$ obtained we choose the interval at whose endpoints the function $f(x)$ assumes values opposite in sign. We designate this interval as $[a_1, b_1]$. Then we again bisect the interval $[a_1, b_1]$ and present the same arguments. We get an interval $[a_2, b_2]$ whose length is $(b - a)/2^2$. We continue the process of bisection until, at some nth stage, either the midpoint of the interval proves to be a root of the equation (a case seldom encountered in practice) or we get an interval $[a_n, b_n]$ such that $b_n - a_n = (b-a)/2^n \leqslant \varepsilon$ and $a_n \leqslant \varepsilon \leqslant b_n$ (the number n indicates how many divisions have been performed). The numbers a_n and b_n are roots of the equation $f(x) = 0$ with an accuracy of ε. As was indicated above, we must take $\xi = (a_n + b_n)/2$ as the approximate value of the root. In this case the error does not exceed $(b - a)/2^{n+1}$.

Example 1. Use the trial and error method to make the smaller root of the equation $x^3 + 3x^2 - 3 = 0$ accurate to $\varepsilon = 10^{-3}$.

△ We use the analytical method to separate the roots of the equation. The function $f(x)$ is defined throughout the number axis. Equating $f'(x)$ to zero, we calculate the root of the derivative:

$$f'(x) = 3x^2 + 6x, \; 3x^2 + 6x = 0, \; x(x + 2) = 0, \; x_1 = 0, \; x_2 = -2.$$

We compile a table of signs of the function:

x	$-\infty$	-2	-1	0	1	$+\infty$
Sign of $f(x)$	$-$	$+$	$-$	$-$	$+$	$+$

We see that the first root lies in the interval $(-\infty, -2)$. We try the root $x = -3$ and find that $f(-3) = -3$:

x	-3	-2	-1	0	1
Sign of $f(x)$	$-$	$+$	$-$	$-$	$+$

This means that the roots of the equation $x^3 + 3x^2 - 3 = 0$ lie in the intervals $(-3, -2)$, $(-2, -1)$, $(0, 1)$.

We refine the smaller root, which lies in the interval $(-3, -2)$, using the bisection method. For the sake of convenience, we compile a table (see Table 5.1). The signs $-$ and $+$ in the upper indices of a_n and b_n mean that $f(a_n) < 0$ and $f(b_n) > 0$.

Thus the root of the equation $x \cong -2.532$. ▲

Table 5.1

n	a_n^-	b_n^+	$x_n=\frac{a_n+b_n}{2}$	x_n^3	$3x_n^2$	$f(x_n)$
0	−3	−2	−2.500	−15.625	18.750	0.125
1	−3	−2.500	−2.750	−20.800	22.689	−1.111
2	−2.750	−2.500	−2.625	−17.990	20.670	−0.320
3	−2.625	−2.500	−2.563	−16.840	19.701	−0.139
4	−2.563	−2.500	−2.532	−16.230	19.233	0.003
5	−2.563	−2.532	−2.548	−16.540	19.479	−0.071
6	−2.548	−2.532	−2.540	−16.390	19.356	−0.034
7	−2.540	−2.532	−2.536	−16.310	19.293	−0.014
8	−2.536	−2.532	−2.534	−16.270	19.263	−0.007
9	−2.534	−2.532	−2.533	−16.250	19.248	−0.002
10	−2.533	−2.532				

Example 2. Use graphical means to separate the root of the equation $x^2 \log_{0.5}(x+1) = 1$. Calculate the root by the method of trial and error with an accuracy of $\varepsilon = 10^{-3}$.

△ We represent the equation in the form $\log_{0.5}(x+1) = 1/x^2$ and construct the graphs of the functions $y = \log_{0.5}(x+1)$

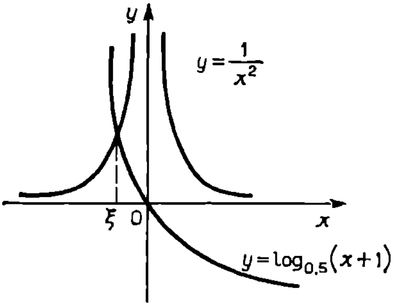


Fig. 5.17

and $y = 1/x^2$. We see from Fig. 5.17 that the equation has one root $x_1 \cong -0.7$. We determine the signs of the function on the left and on the right of x_1:

x	−0.8	−0.5
Sign of $f(x)$	+	−

To make the calculations more convenient, we pass to the decimal logarithms:

$$f(x)=x^2\frac{\log(x+1)}{\log 0.5}-1=x^2\frac{\log(x+1)}{-0.301}-1.$$

We compile the following table (Table 5.2):

Table 5.2

n	a_n^+	b_n^-	$x_n=\frac{a_n+b_n}{2}$	x_n^2	$\log(x_n+1)$	$f(x_n)$
0	−0.8	−0.5	−0.65	0.4225	−0.4559	−0.360
1	−0.8	−0.65	−0.73	0.5329	−0.5686	−0.0067
2	−0.73	−0.65	−0.69	0.4761	−0.5086	−0.196
3	−0.73	−0.69	−0.71	0.5041	−0.5376	−0.099
4	−0.73	−0.71	−0.72	0.5184	−0.5528	−0.048
5	−0.73	−0.72				

Thus $x_1 \cong -0.73$. ▲

5.4. Method of Chords

The **method of chords** is one of the most widely used methods of solving algebraic and transcendental equations. In literature it is also encountered under the names of "the method of false position" (regula falsi method) and "the method of linear interpolation".

Consider an equation $f(x) = 0$, where $f(x)$ is a continuous function which has derivatives of the first and the second order in the interval (a, b). The root is assumed to be separated and is on the interval $[a, b]$, i.e. $f(a)\cdot f(b) < 0$.

The idea of this method is that on a sufficiently small interval $[a, b]$ the arc of the curve $y = f(x)$ is replaced by the chord subtending it. The point of intersection of the chord and the x-axis is taken as the approximate value of the root.

We have considered earlier various variants of the position of the arc of the curve depending on the signs of the first and the second derivative.

1*st case*. The first and the second derivative are of the same sign, i.e. $f'(x)\cdot f''(x) > 0$. Assume, for instance, that $f(a) < 0$, $f(b) > 0$, $f'(x) > 0$ and $f''(x) > 0$

(Fig. 5.18 *a*). The graph of the function passes through the points A_0 $(a,\ f(a))$, B $(b,\ f(b))$. The required root of the equation $f(x) = 0$ is the abscissa of the point of intersection of the graph of the function $y = f(x)$ and x-axis. We do not know that point and, instead of it, take the point x_1 of intersection of the chord A_0B and the x-axis. And this is the approximate value of the root.

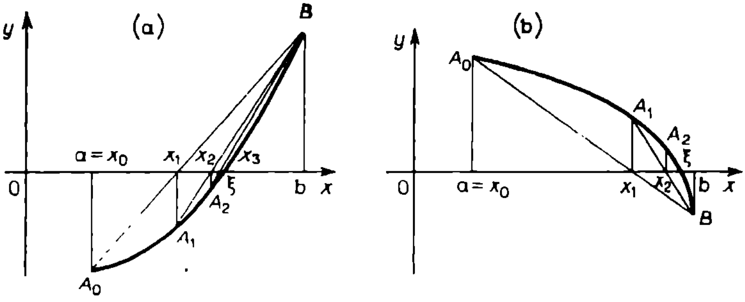


Fig. 5.18

The equation of the chord which passes through the points A_0 and B has the form

$$\frac{y - f(a)}{f(b) - f(a)} = \frac{x - a}{b - a}.$$

We seek the value of $x = x_1$ for which $y = 0$:

$$x_1 = a - \frac{f(a)(b-a)}{f(b) - f(a)}. \tag{1}$$

The root ξ is now within the interval $[x_1,\ b]$. If the value of the root x_1 does not suit our purpose, we can refine it, applying the method of chords to the interval $[x_1,\ b]$. We connect the point A_1 $(x_1,\ f(x_1))$ with the point B $(b,\ f(b))$ and find x_2, which is the point of intersection of the chord A_1B and the x-axis:

$$x_2 = x_1 - \frac{f(x_1)(b - x_1)}{f(b) - f(x_1)}.$$

Continuing this process, we obtain

$$x_3 = x_2 - \frac{f(x_2)(b-x_2)}{f(b)-f(x_2)}$$

and, in general,

$$x_{n+1} = x_n - \frac{f(x_n)(b-x_n)}{f(b)-f(x_n)}. \tag{2}$$

The process goes on until we get the approximate root with the specified degree of accuracy.

The formulas given above are also used to calculate roots when $f(a) > 0$, $f(b) < 0$, $f'(x) < 0$, $f''(x) < 0$ (Fig. 5.18 *b*).

2nd case. The first and the second derivative are of unlike signs, i.e. $f'(x) \cdot f''(x) < 0$. Assume, for instance, that $f(a) > 0$, $f(b) < 0$, $f'(x) < 0$, $f''(x) > 0$ (Fig. 5.19 *a*). We connect the points $A(a, f(a))$ and $B_0(b, f(b))$ and write the equation of the chord which passes through the points A and B_0:

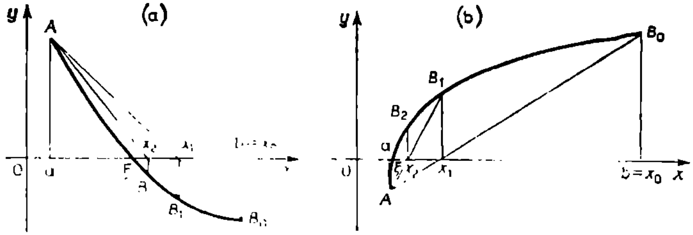


Fig. 5.19

$$\frac{y-f(b)}{f(b)-f(a)} = \frac{x-b}{b-a}.$$

We seek x_1 as the point of intersection of the chord and the x-axis, setting $y = 0$:

$$x_1 = b - \frac{f(b)(b-a)}{f(b)-f(a)}. \tag{3}$$

The root ξ is now within the interval $[a, x_1]$. Applying the method of chords to the interval $[a, x_1]$, we get

$$x_2 = x_1 - \frac{f(x_1)(x_1-a)}{f(x_1)-f(a)}$$

and, in general,

$$x_{n+1} = x_n - \frac{f(x_n)(x_n - a)}{f(x_n) - f(a)}. \tag{4}$$

Using these formulas, we can also find the approximate value of the root when $f(a) < 0$, $f(b) > 0$, $f'(x) > 0$, $f''(x) < 0$ (Fig. 5.19 *b*).

Thus, if $f'(x) \cdot f''(x) > 0$, then the approximate root can be found from formulas (1) and (2) and if $f'(x) \times f''(x) < 0$, then it can be found from formulas (3) and (4).

However, we can choose the convenient formulas employing the following simple rule: *the stationary endpoint of an interval is the endpoint for which the sign of the function coincides with the sign of the second derivative.*

If $f(b) \cdot f''(x) > 0$, then the endpoint b is stationary and we can take the endpoint a as the initial approximation [formulas (1) and (2)]. Now if $f(a) \cdot f''(x) > 0$, then the endpoint a is stationary and we must take the endpoint b as the initial approximation [formulas (3) and (4)].

Thus, as a result of the repeated application of formulas (2) and (4), we get a monotonic sequence $x_1, x_2, x_3, \ldots, x_n$ which converges to the value of the root ξ.

To estimate the error of approximation, we can use the formula

$$|\xi - x_n| < |x_n - x_{n-1}|, \tag{5}$$

where ξ is the exact value of the root and x_{n-1} and x_n are its approximations obtained at the $(n-1)$th and nth stages. It can be used when the following condition is fulfilled:

$$M \leqslant 2m, \text{ where } M = \max_{[a,\,b]} |f'(x)|, \quad m = \min_{[a,\,b]} |f'(x)|. \tag{6}$$

Example 1. Use the method of chords to make the smaller root of the equation $x^3 + 3x^2 - 3 = 0$ accurate to $\varepsilon = 0.001$. The roots of the equation have been separated and the smaller root is on the interval $[-3, -2]$ (see Example 1 in 5.3).

△ We verify the fulfillment of condition (6):

$$|f'(x)| = |3x^2 + 6x|,$$

$$M = \max_{[-3,\,-2]} |f'(x)| = |27 - 18| = 9,$$

$$m = \min_{[-3,\,-2]} |f'(x)| = |12 - 12| = 0, \quad M > 2m.$$

We take the midpoint of the interval $[-3, -2]$, i.e. the point $x = -2.5$, and choose the interval $[-3, -2.5]$. We again verify the fulfillment of condition (6):

$$M = \max_{[-3,\ -2.5]} |f'(x)| = 9, \quad m = \min_{[-3,\ -2.5]} |f'(x)| = 3.75,$$
$$M > 2m.$$

Now we take the midpoint of the interval $[-3, -2.5]$, i.e. the point $x = -2.75$; we have $f(-2.75) < 0$, $f(-2.5) > 0$, $f(-3) < 0$. We choose the interval $[-2.75, -2.5]$ and find that

$$M = \max_{[-2.75,\ -2.5]} |f'(x)| = 6.189, \quad m = \max_{[-2.75,\ -2.5]} |f'(x)| = 3.75,$$

i.e. in this case the condition $M < 2m$ is fulfilled.

Thus, to estimate the error of the root lying on the interval $[-2.75, -2.5]$, we can use formula (5): $|\xi - x_n| < |x_n - x_{n-1}|$, i.e. we must continue the process of the successive approximation of the root until the condition $|x_n - x_{n-1}| \leqslant \varepsilon$ is fulfilled.

We determine the sign of the second derivative and find the formula which must be used for calculations. We find that $f''(x) = 6x + 6$. The inequalities $f(-2.75) < 0$ and $f(-2.75)\cdot f''(x) > 0$ hold true on the interval $[-2.75, -2.5]$. This means that we must take $x = -2.75$ as the stationary endpoint of the interval. Then we must use formulas (3) and (4) to carry out the calculations:

$$x_1 = b - \frac{f(a)(b-a)}{f(b)-f(a)}, \quad x_{n+1} = x_n - \frac{f(x_n)(x_n-a)}{f(x_n)-f(a)},$$

where $a = -2.75$ and $f(a) = -1.111$. If we represent the last expression as

$$x_{n+1} - x_n = -\frac{f(x_n)(x_n-a)}{f(x_n)-f(a)},$$

we can at once get the difference of the two successive approximations and verify whether we can terminate the calculations, i.e. whether the inequality $|x_{n+1} - x_n| \leqslant \varepsilon$ is satisfied.

It is convenient to use the following table to carry out the calculations:

Table 5.3

n	x_n	x_n^3	x_n^2	$3x_n^2$	$f(x_n) = x_n^3 + 3x_n^2 - 3$	$x_n - a$	$-\frac{f(x)(x-a)}{f(x_n)-f(a)}$
0	−2.5	−15.625	6.250	18.75	0.125	0.25	−0.025
1	−2.525	−16.098	6.3756	19.1268	0.0288	0.225	−0.006
2	−2.531	−16.213	6.4060	19.2180	0.0050	0.219	−0.0009
3	−2.5319						

We can see from Table 5.3 that $|x_3 - x_2| < 0.001$ and, therefore, rounding off to the thousands place, we get $\xi = -2.532$. ▲

Example 2. Use the method of chords to make the root of the equation $x - \sin x = 0.25$, which is on the interval $[0, \pi/2]$, accurate to $\varepsilon = 0.001$.

△ We write the equation in the form $x - \sin x - 0.25 = 0$ and find $f'(x) = 1 - \cos x$. To verify whether condition (6) is fulfilled, we compile an auxiliary table whose first two columns indicate the origin and the endpoint of the chosen interval of the separation of the root.

Table 5.4

a	b	Signs of		M	m	Fulfillment of the condition $M \leqslant 2m$	$\frac{a+b}{2}$	Sign of $f\left(\frac{a+b}{2}\right)$
		$f(a)$	$f(b)$					
0.00	1.57	−	+	1.00		no	0.785	−
0.785	1.57	−	+	1.00	0.2925	no	1.178	+
0.785	1.178	−	+	0.6172	0.2925	no	0.982	−
0.982	1.178	−	+	0.6172	0.4446	yes		

We see from the last row of the table that on the interval [0.982, 1.178] the condition $M \leqslant 2m$ is fulfilled. Consequently, when using the method of chords to estimate the error of the approximation of the root, we can employ the inequality $|x_n - x_{n-1}| < \varepsilon$. The root of the equation $x - \sin x - 0.25 = 0$ is on the interval [0.982, 1.178]. We determine the sign of the second derivative within the interval:

$$f'(x) = 1 - \cos x,\quad f''(x) = \sin x > 0.$$

If we return to the old designations, then $a = 0.982$ and $b = 1.178$. The sign of the second derivative coincides with the sign of the function at the point b. Consequently, this endpoint of the interval is stationary and all the approximations of the root are from the side of the endpoint a. We use formulas (1) and (2) to calculate the root:

$$x_1 = a - \frac{f(a)(b-a)}{f(b)-f(a)}; \quad x_{n+1} = x_n - \frac{f(x_n)(b-x_n)}{f(b)-f(x_n)},$$

where $b = 1.178$, $f(b) = 0.00416$. We compile the following table:

Table 5.5

n	x_n	$-\sin x_n$	$f(x_n) = x_n - \sin x_n - 0.25$	$b - x_n$	$-\frac{f(x_n)(b-x_n)}{f(b)-f(x_n)}$
0	0.982	−0.83161	−0.09961	0.196	0.189
1	1.171	−0.92114	−0.00014	0.007	0.0002
2	1.1712				

Thus $x = 1.171$ with an accuracy of $\varepsilon = 0.001$. ▲

5.5. Newton's Method of Approximation

Newton's method is another method of iteration.

Assume that the root of the equation $f(x) = 0$ has been separated on the interval $[a, b]$ and $f'(x)$ and $f''(x)$ are continuous and retain constant signs throughout the interval $[a, b]$.

In terms of geometry, the meaning of Newton's method is that the arc of the curve $y = f(x)$ is replaced by a tangent to that curve (hence, this method is sometimes called the **method of tangents**).

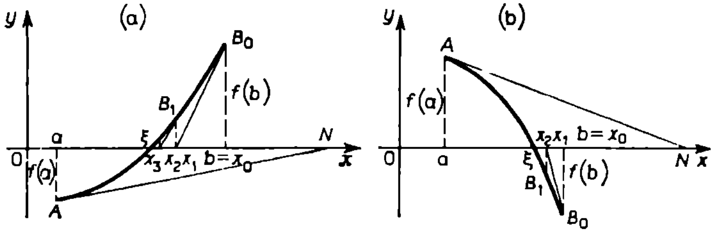


Fig. 5.20

1st case. The first and the second derivative have the same sign. Let $f(a) < 0$, $f(b) > 0$, $f'(x) > 0$, $f''(x) > 0$ (Fig. 5.20 *a*) or $f(a) > 0$, $f(b) < 0$, $f'(x) < 0$, $f''(x) < 0$ (Fig. 5.20 *b*). We draw a tangent to the curve $y = f(x)$ at the point $B_0(b, f(b))$ and find the abscissa of the point of intersection of the tangent and the x-axis. We know that the equation of the tangent at the point $B_0(b, f(b))$ has the form

$$y - f(b) = f'(b)(x - b).$$

Setting $y = 0$ and $x = x_1$, we obtain

$$x_1 = b - \frac{f(b)}{f'(b)}. \tag{1}$$

The root of the equation is now on the interval $[a, x_1]$. Using again Newton's method, we draw a tangent to the curve at the point $B_1(x_1, f(x_1))$ and obtain

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$

and, in general,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{2}$$

We get a sequence of approximate values x_1, x_2, . . ., x_n, . . ., every successive term of which is closer to the root ξ than its predecessor. However, all x_n remain larger than the exact root ξ, i.e. x_n is an approximate value of the root ξ by excess.

2nd case. The first and the second derivative are of unlike signs. Let $f(a) < 0$, $f(b) > 0$, $f'(x) > 0$, $f''(x) < 0$ (Fig. 5.21 *a*) or $f(a) > 0$, $f(b) < 0$, $f'(x) < 0$, $f''(x) > 0$ (Fig. 5.21 *b*). If we again draw a tangent to the curve $y = f(x)$ at the point B, it will cut the abscissa axis at the point which does not belong to the interval $[a, b]$. We therefore draw a tangent at the point A_0 $(a, f(a))$ and write its equation for this case:

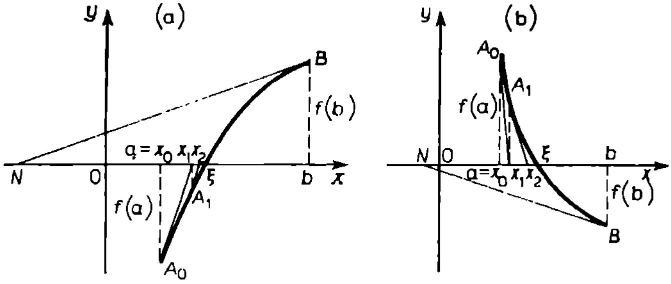


Fig. 5.21

$$y - f(a) = f'(a)(x - a).$$

Setting $y = 0$ and $x = x_1$, we find that

$$x_1 = a - \frac{f(a)}{f'(a)}. \tag{3}$$

The root ξ is now on the interval $[x_1, b]$. Using again Newton's method, we draw a tangent at the point A_1 $(x_1, f(x_1))$ and get

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$

and, in general,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{4}$$

We get a sequence of approximate values $x_1, x_2, \ldots, x_n, \ldots$, each successive term of which is closer to the exact value of the root ξ than its predecessor, i.e. x_n is an approximate value of the root ξ by defect.

Comparing these formulas with those derived earlier, we note that they differ only by the choice of the initial approximation: in the first case we assumed the endpoint b of the interval to be x_0 and in the second case, the endpoint a.

When choosing the initial approximation of a root we must be guided by the following **rule**: *the endpoint of the interval $[a, b]$ at which the sign of the function coincides with the sign of the second derivative must be taken as the initial point.* In the first case $f(b) \cdot f''(x) > 0$ and the initial point $b = x_0$, in the second case $f(a) \cdot f''(x) > 0$ and we take $a = x_0$ as the initial approximation.

To estimate the error, we can use the general formula

$$|\xi - x_n| \leqslant \frac{|f(x_n)|}{m}, \quad \text{where } m = \min_{[a,\, b]} |f'(x)| \tag{5}$$

(this formula can also be used for the method of chords).

When the interval $[a, b]$ is so small that the condition $M_2 < 2m_1$, where $M_2 = \max\limits_{[a,\, b]} |f''(x)|$ and $m_1 = \min\limits_{[a,\, b]} |f'(x)|$, is fulfilled on it, the accuracy of approximation at the nth stage is estimated as follows: if

$$|x_n - x_{n-1}| < \varepsilon \text{ then } |\xi - x_n| < \varepsilon^2.$$

If the derivative $f'(x)$ varies but slightly on the interval $[a, b]$, then, to simplify the calculations, we can use the formula

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)}, \tag{6}$$

i.e. it is sufficient to calculate the value of the derivative at the origin only once. In terms of geometry this means that the tangents at the points B_n $(x_n, f(x_n))$ are replaced by straight lines which are parallel to the tangent drawn to the curve $y = f(x)$ at the point B_0 $(x_0, f(x_0))$ (Fig. 5.22).

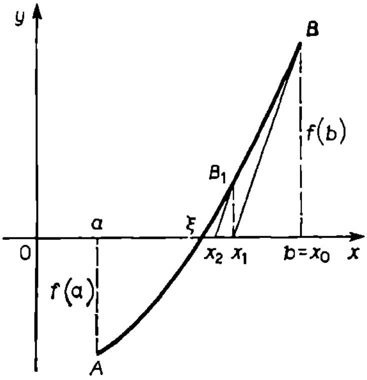


Fig. 5.22

Example 1. Use Newton's method to find the root of the equation $x^3 + 3x^2 - 3 = 0$, which lies on the interval $[-2.75, -2.5]$ with an accuracy of $\varepsilon = 0.001$.

△ We have established the fact that $f(-2.75) \cdot f''(x) > 0$ (see Example 1 in 5.4). Therefore, to use the method of tangents, we must choose $x_0 = -2.75$. We shall use formula (6) to carry out the calculations. We find that

$$f'(x) = 3x^2 + 6x, \quad f'(x_0) = f'(-2.75) = 6.1875.$$

For the sake of convenience, we shall use Table 5.6, from which we can see that $|x_5 - x_4| < 0.001$, and, therefore, ξ — 2.533. ▲

Table 5.6

n	x_n	x_n^3	x_n^2	$3x_n^2$	$f(x_n)$	$-\dfrac{f(x_n)}{6.1875}$
0	−2.75	−20.797	7.5625	22.6875	−1.111	0.179
1	−2.571	−16.994	6.6100	19.8300	−0.164	0.026
2	−2.545	−16.484	6.4770	19.431	−0.053	0.008
3	−2.537	−16.329	6.4364	19.309	0.020	0.003
4	−2.534	−16.271	6.4212	19.2636	0.007	0.001
5	−2.533					

Example 2. Use Newton's method to find the root of the equation $x - \sin x = 0.25$, which lies on the interval $[0.982, 1.178]$, with an accuracy of $\varepsilon = 0.0001$.

△ Here we have $a = 0.982$, $b = 1.178$. We find that $f'(x) = 1 - \cos x$, $f''(x) = \sin x > 0$ on $[0.982, 1.178]$, $f(1.178) \cdot f''(x) > 0$. Consequently, $x_0 = 1.178$. We use formulas (1) and (2) and Table 5.7 to carry out the calculations. We can see from Table 5.7 that $|x_3 - x_2| < 0.0001$. Thus $\xi \simeq 1.1712$. ▲

Table 5.7

n	x_n	$-\sin x_n$	$f(x_n) = x_n - \sin x_n - 0.25$	$f'(x_n) = 1 - \cos x_n$	$-\dfrac{f(x_n)}{f'(x_n)}$
0	1.178	−0.92384	0.00416	0.61723	−0.0065
1	1.1715	−0.92133	0.00017	0.61123	−0.0002
2	1.1713	−0.92127	0.00003	0.61110	−0.00005
3	1.17125				

5.6. The Combination of the Method of Chords and Newton's Method

The method of chords and Newton's method yield approximations of a root from different sides and therefore are often used in combination. Then the accuracy of the root increases much quicker.

Assume that we have an equation $f(x) = 0$. The root ξ has been separated and is on the interval $[a, b]$. Taking into account the type of the graph of the function we employ the **combination of the method of chords and Newton's approximation method.**

If $f'(x) \cdot f''(x) > 0$, *then the method of chords yields approximations of the root by defect and Newton's method yields approximations by excess* (Fig. 5.23 *a* and *b*).

Now if $f'(x) \cdot f''(x) < 0$, *we can use the method of chords to get the values of the root by excess and Newton's method to get the values by defect* (Fig. 5.24 *a*, *b*).

However, in all cases the true value of the root is between the approximate values of the roots obtained by the method of chords and Newton's method, i.e. there holds an inequality $a < \bar{x}_n < \xi < \bar{\bar{x}}_n < b$, where $\bar{x}_n$ is the approximate value of the root by defect and $\bar{\bar{x}}_n$ by excess.

The calculations must be carried out as follows. If $f'(x) \cdot f''(x) > 0$, then we must take the endpoint a as

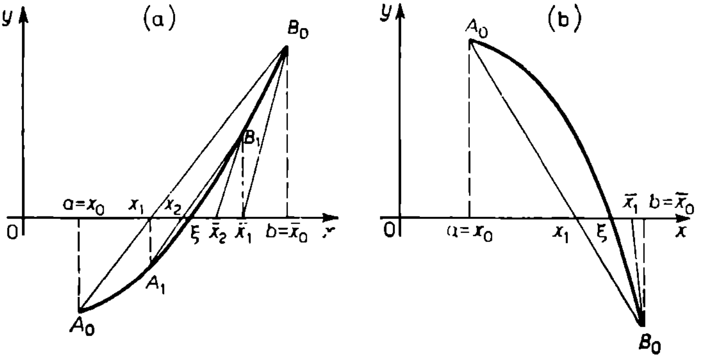


Fig. 5.23

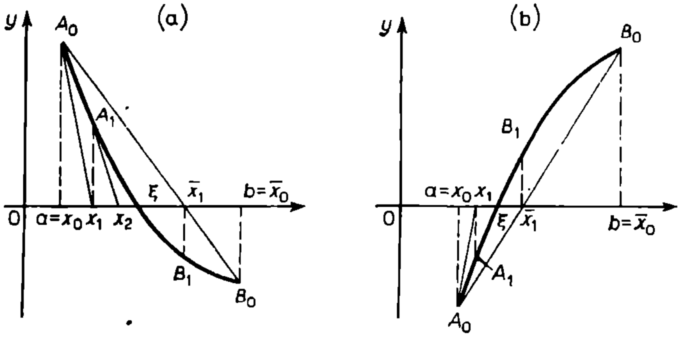


Fig. 5.24

the initial approximation for the method of chords and the endpoint b for Newton's method and then

$$a_1 = a - \frac{f(a)(b-a)}{f(b)-f}, \quad b_1 = b - \frac{f(b)}{f'(b)}. \qquad (1)$$

The true value of the root is now on the interval $[a_1, b_1]$. Applying the combined method to this interval, we obtain $a_2 = a_1 - \frac{f(a_1)(b_1-a_1)}{f(b_1)-f(a_1)}$, $b_2 = b_1 - \frac{f(b_1)}{f'(b_1)}$

and, in general,

$$a_n \qquad \frac{f(a_n)(b_n-a_n)}{f(b_n)-f(a_n)}, \quad b_{n+1} = b_n - \frac{f(b_n)}{f'(b_n)} \qquad (2)$$

(see Fig. 5.23 *a*, *b*).

Now if $f'(x) \cdot f''(x) < 0$, then we must take the endpoint b as the initial approximation for the method of chords and the endpoint a for Newton's method. Then we have

$$a_1 = a - \frac{f(a)}{f'(a)}, \quad b_1 = b - \frac{f(b)(b-a)}{f(b) - f(a)}. \qquad (3)$$

Applying the combined method to the interval $[a_1, b_1]$ we got

$$a_2 = a_1 - \frac{f(a_1)}{f'(a_1)}, \quad b_2 = b_1 - \frac{f(b_1)(b_1 - a_1)}{f(b_1) - f(a_1)}$$

and, in general,

$$a_{n+1} = a_n - \frac{f(a_n)}{f'(a_n)}, \quad b_{n+1} = b_n - \frac{f(b_n)(b_n - a_n)}{f(b_n) - f(a_n)} \qquad (4)$$

(see Fig. 5.24 *a*, *b*).

The combined method is very convenient for the evaluation of the error of calculations. The process of calculations is terminated as soon as the inequality $|\bar{\bar{x}}_n - \bar{x}_n| < \varepsilon$ is satisfied. We must take

$$\xi = \frac{1}{2}(\bar{x}_n + \bar{\bar{x}}_n), \qquad (5)$$

where $\bar{x}_n$ and $\bar{\bar{x}}_n$ are the approximate values of the root by defect and by excess respectively, as the approximate value of the root.

Example. Use the combination of the method of chords and Newton's method to find the roots of the equation $x^3 + 3x^2 - 24x + 1 = 0$ with the accuracy of 0.001.

△ (1) We separate the roots analytically, and have

$$f(x) = x^3 + 3x^2 - 24x + 1, \quad f'(x) = 3x^2 + 6x - 24,$$

i.e. the roots of the derivative are $x_1 = -4$ and $x_2 = 2$. We compile a table of signs of the function:

x	$-\infty$	-4	2	$+\infty$
Sign of $f(x)$	$-$	$+$	$-$	$+$

The equation has three real roots: $x_1 \in (-\infty, -4)$, $x_2 \in (-4, 2)$, $x_3 \in (2, +\infty)$. We diminish the intervals of seeking the roots, making them equal to 1:

x	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4
Sign of $f(x)$	$-$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$-$	$-$	$-$	$+$

Thus, $x_1 \in (-7, -6)$, $x_2 \in (0, 1)$, $x_3 \in (3, 4)$.

(2) We use the combination of the method of chords and Newton's method to make the root lying in the interval $(-7, -6)$ more precise and get $f(-7) = -27 < 0$, $f(-6) = 37 > 0$, $f'(x) = 3x^2 + 6x - 24 > 0$, $f''(x) = 6x + 6 < 0$, $f'(x) \cdot f''(x) < 0$.

We calculate employing formulas (4):

$$a_{n+1} = a_n - \frac{f(a_n)}{f'(a_n)}, \quad b_{n+1} = b_n - \frac{f(b_n)(b_n - a_n)}{f(b_n) - f(a_n)},$$

i.e.

$$a_{n+1} = a_n + \Delta a_n, \quad \text{where} \quad \Delta a_n = -\frac{f(a_n)}{f'(a_n)},$$

$$b_{n+1} = b_n + \Delta b_n, \quad \text{where} \quad \Delta b_n = -\frac{f(b_n)(b_n - a_n)}{f(b_n) - f(a_n)}$$

(a_n and b_n are the approximate values of the root by defect and by excess respectively). Here $a_0 = a = -7$, $b_0 = b = -6$.

It is convenient to carry out the calculations using a table (see Table 5.8).

Taking into account that $|b_2 - a_2| = 0.0007 < 0.001$, we must terminate the calculations and take

$$\xi_1 = \frac{1}{2}(-6.6384 - 6.6377) = -6.638$$

as the approximate value of the root ξ_1.

(3) Let us estimate the approximate value of the root for the interval $(0, 1)$. We have $f(0) > 0$, $f(1) < 0$, $f'(x) = 3x^2 + 6x - 24 < 0$, $f''(x) = 6x + 6 > 0$, $f'(x) \cdot f''(x) < 0$. As in the first case, we use formulas (4) for $a_0 = a = 0$, $b_0 = b = 1$.

We compile a table (see Table 5.9). Thus we can take $\xi_2 = 0.042$ with an accuracy of 0.001.

(4) Let us now find the approximate value of the root belonging to the interval $(3, 4)$. We have $f(3) = -17 < 0$, $f(4) = 17 > 0$, $f'(x) = 3x^2 + 6x - 24 > 0$, $f''(x) = 6x + 6 > 0$, $f'(x) \cdot f''(x) > 0$.

We use formulas (2) to carry out the calculations:

$$a_{n+1} = a_n - \frac{f(a_n) \cdot (b_n - a_n)}{f(b_n) - f(a_n)}, \quad b_{n+1} = b_n - \frac{f(b_n)}{f'(b_n)},$$

i.e.

$$a_{n+1} = a_n + \Delta a_n, \quad \text{where} \quad \Delta a_n = -\frac{f(a_n)(b_n - a_n)}{f(b_n) - f(a_n)},$$

$$b_{n+1} = b_n + \Delta b_n, \quad \text{where} \quad \Delta b_n = -\frac{f(b_n)}{f'(b_n)}.$$

Here $a_0 = a = 3$, $b_0 = b = 4$.

We reduce the calculations to a table (see Table 5.10).

Thus we can take $\xi_3 = 3.596$ with an accuracy of 0.001. ▲

Table 5.8

n	a_n	b_n-a_n	a_n^2	a_n^3	$f(a_n)$	$f'\ a_n)$	$f(b_n)-f(a_n)$	Δa_n	a_{n-1}
	b_n		b_n^2	b_n^3	$f(b_n)$			Δb_n	b_{n+1}
0	−7	1	49	−343	−27	81	64	0.3333	−6.6667
	−6		36	−216	37			−0.5781	−6.5781
1	−6.6667	0.0886	44.4449	−296.3007	−1.9652	69.3345	6.01002	0.0283	−6.6384
	−6.5781		43.2714	−284.6436	4.0450			−0.0596	−6.6377
2	−6.6384	0.0007							
	−0.6377								

Table 5.9

n	a_n	b_n-a_n	a_n^2	a_n^3	$f(a_n)$	$f'(a_n)$	$f(b_n)-f(a_n)$	Δa_n	a_{n+1}
	b_n		b_n^2	b_n^3	$f(b_n)$			Δb_n	b_{n+1}
0	0	1	0	0	1	−24	−20	0.0417	0.0417
	1		1	1	−19			−0.9500	0.0500
1	0.0417	0.0083	0.0017	0.00007	0.0045	−23.7446	−0.1969	0.0002	0.0419
	0.0500		0.0025	0.0001	−0.1924			−0.0081	0.0419
2	0.0419	0.0000							
	0.0419								

Table 5.10

n	a_n	b_n-a_n	a_n^2	a_n^3	$f(a_n)$	$f'(b_n)$	$f(b_n)-f(a_n)$	Δa_n	a_{n+1}
	b_n		b_n^2	b_n^3	$f(b_n)$			Δb_n	b_{n+1}
0	3	1	9	27	−17	48	34	0.5000	3.5000
	4		16	64	17			−0.3542	3.6458
1	3.5000	0.1458	12.2500	42.875	−3.3750	37.7505	5.2110	0.0944	3.5944
	3.6458		13.2919	48.4595	1.8360			−0.0486	3.5972
2	3.5944	0.0028	12.9197	46.4386	−0.0679	36.4020	0.1017	0.0019	3.5963
	3.5972		12.9398	46.5472	0.0338			−0.0009	3.5963
3	3.5963	0.0000							
	3.5963								

5.7. The Iterative Method

The gist of the method. The **iterative method**, or the **method of successive approximations**, is one of the most important methods in computational mathematics. The main advantage of this method is that the operations carried out at each stage are of the same kind, and this makes it considerably easier to set up programs for a computer which are based on iterative algorithms.

The gist of the iterative method is the following. Consider the equation

$$f(x) = 0. \tag{1}$$

Let $f(x)$ be a function, continuous on the interval $[a, b]$ which vanishes within this interval at least at one point ξ. We have to find at least one of its real roots, which lie on $[a, b]$, with a specified accuracy.

We replace equation (1) by an equivalent equation, i.e. by an equation which has the same roots, say, by an equation of the form

$$x = \varphi(x). \tag{2}$$

We choose some value $x_0 \in [a, b]$, say, $x_0 = (a + b)/2$, as the initial approximation. Then we calculate $\varphi(x_0)$ and assume the resultant number $x_1 = \varphi(x_0)$ to be the first approximation of the value of the root ξ. Substituting x_1 for x on the right-hand side of equation (2), we get a new number $x_2 = \varphi(x_1)$. Continuing this procedure, we arrive at a sequence of numbers $x_0, x_1, x_2, \ldots, x_n, \ldots$, which are defined by the following relations:

$$x_0 = (a + b)/2, \quad x_n = \varphi(x_{n-1}) \quad (n = 1, 2, \ldots). \tag{3}$$

If there is a limit $\lim\limits_{n\to\infty} x_n = \xi$ and the function $\varphi(x)$ is continuous, then we can pass to the limit in relation (3) and obtain

$$\xi = \varphi(\xi), \tag{4}$$

i.e. the limit ξ is a root of equation (2) and, consequently, of equation (1) as well. Since process (3) is convergent, for a sufficiently large n this root can be calculated with any specified accuracy.

Note that there are infinitely many ways of arriving at representation (2) [i.e. at the form of the function $\varphi(x)$]. This is very significant since the form of the function $\varphi(x)$ is of a considerable importance both for the convergence itself and for its rate (provided that the fact of convergence has been established).

The following theorem defines the conditions for convergence of the iterative process (3).

Theorem. *Assume that the following conditions are fulfilled:*

(1°) *the function $\varphi(x)$ is defined and differentiable on the interval $[a, b]$,*

(2°) *all the values of $\varphi(x) \in [a, b]$ for $x \in [a, b]$,*

(3°) *there is a number $q < 1$ such that*

$$|\varphi'(x)| \leqslant q < 1 \tag{5}$$

for $x \in [a, b]$.

Then the iterative process (3) *converges irrespective of the choice of the initial approximation $x_0 \in [a, b]$ and $\lim\limits_{n\to\infty} x_n$ is the unique and simple root of the equation $x = \varphi(x)$ on the interval $[a, b]$.*

☐ We set up the following differences:

$$\begin{aligned}
&|x_1 - x_0| = |x_1 - x_0|,\\
&|x_2 - x_1| = |\varphi(x_1) - \varphi(x_0)| = |\varphi'(c_1)| \cdot |x_1 - x_0|\\
&\qquad \leqslant q|x_1 - x_0|,\\
&|x_3 - x_2| = |\varphi(x_2) - \varphi(x_1)| = |\varphi'(c_2)| \cdot |x_2 - x_1|\\
&\qquad \leqslant q^2|x_1 - x_0|,\\
&\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\\
&|x_n - x_{n-1}| = |\varphi(x_{n-1}) - \varphi(x_{n-2})|\\
&\qquad |\varphi'(c_n)| \cdot |x_{n-1} - x_{n-2}| \leqslant q^{n-1}|x_1 - x_0|.
\end{aligned}$$

Here $c_k \in [x_{k-1}, x_k]$ and $x_k \in [a, b]$ by virtue of condition 2°.

Consider a series with the following partial sums:

$$S_{n+1} = x_0 + (x_1 - x_0) + (x_2 - x_1) + \dots + (x_n - x_{n-1} \dots) + \dots .$$

It is evident that $S_{n+1} = x_n$. The series $x_0 + \sum_{i=0}^{\infty} (x_{i+1} - x_i)$ is convergent since all of its terms beginning with $x_1 - x_0$, do not exceed, in absolute value, the terms of the geometric progression with a common ratio $q < 1$. Hence there is a limit

$$\lim_{n\to\infty} S_{n+1} = \lim_{n\to\infty} x_n = \xi.$$

Since the function $\varphi(x)$ is continuous, we have $\lim_{n\to\infty} \varphi(x_{n-1}) = \varphi(\xi)$ and since $x_n = \varphi(x_{n-1})$, it follows that $\varphi(\xi) = \xi$, i.e. $\xi = \lim_{n\to\infty} x_n$ is a root of equation (2).

Let us prove that this root is unique. Let ξ_1 and ξ_2 be two roots of equation (2), i.e. $\xi_1 = \varphi(\xi_1)$ and $\xi_2 = \varphi(\xi_2)$. Then

$$\begin{aligned} |\xi_1 - \xi_2| &= |\varphi(\xi_1) - \varphi(\xi_2)| \\ &= |\varphi'(c)| \cdot |\xi_1 - \xi_2|, \end{aligned} \tag{6}$$

where $c \in (\xi_1, \xi_2)$. We reduce relation (6) to the form

$$|\xi_1 - \xi_2| (1 - |\varphi'(c)|) = 0$$

and then it follows from condition 3° that $\xi_1 = \xi_2$, i.e. the two roots are not distinct.

Let us finally prove that the root obtained is simple. To do this, it is sufficient to prove that $x - \varphi(x)$ has a derivative which does not vanish at any point of the interval $[a, b]$. Indeed, $(x - \varphi(x))' = 1 - \varphi'(x)$, and it is evident, by virtue of condition 3°, that this expression is positive on $[a, b]$. ■

The estimate of the error. Consider the difference of the exact and the approximate value of the root ξ:

$$\begin{aligned} |\xi - x_n| &= |\varphi(\xi) - \varphi(x_{n-1})| \leqslant q|\xi - x_{n-1}| \\ &= q|\xi - x_n + x_n - x_{n-1}| \leqslant q|\xi - x_n| + q|x_n - x_{n-1}|. \end{aligned}$$

From this we have

$$|\xi - x_n| \leqslant \frac{q}{1-q} |x_n - x_{n-1}| \leqslant \frac{q^n}{1-q} |x_1 - x_0|. \tag{7}$$

Relation (7) makes it possible to find, already after the first iteration, the maximum number of iterations $N(\xi)$, necessary to calculate the root with the specified accuracy ε. Indeed, for $|\xi - x_n|$ to be not larger than ε, it is sufficient that

$$\frac{q^n}{1-q}|x_1 - x_0| \leqslant \varepsilon,$$

whence we have

$$N(\varepsilon) \geqslant \frac{\log \frac{\varepsilon(1-q)}{|x_1 - x_0|}}{\log q}. \tag{8}$$

For $q \leqslant 1/2$, the estimate of the error becomes simpler and assumes the form

$$|\xi - x_n| \leqslant |x_n - x_{n-1}|. \tag{9}$$

We have mentioned that the form of the equation $x = \varphi(x)$ is of importance for the convergence of the iterative method. We shall now show a sufficiently general technique for constructing the function $\varphi(x)$, for which the fulfillment of condition 3° of the theorem is ensured.

Let us consider the initial equation $f(x) = 0$. Assume that there is a unique root ξ of the equation on the interval $[a, b]$ and that the derivative $f'(x)$ exists for $x \in [a, b]$ and retains sign so that

$$m_1 \leqslant f'(x) \leqslant M_1, \quad \text{where} \quad M_1 = \max_{[a,\, b]} f'(x);$$
$$m_1 = \min_{[a,\, b]} f'(x) \tag{10}$$

(without loosing generality, we can assume that $f'(x) > 0$).

We replace the equation $f(x) = 0$ by an equivalent equation

$$x = x - \lambda f(x) \tag{11}$$

and choose the constant λ which would ensure the fulfillment of condition 3°.

For the function $\varphi(x) \equiv x - \lambda f(x)$ condition 3° is written as follows:

$$|1 - \lambda f'(x)| < 1$$

We solve this inequality and have $-1 < 1 - \lambda f'(x) < 1$. From the right-hand inequality we find that $\lambda > 0$

and from the left-hand one we get $\lambda < 2/f'(x)$, i.e.

$$0 < \lambda < 2/f'(x), \quad x \in [a, b],$$

or

$$0 < \lambda < 2/M_1.$$

We usually assume $1/M_1$ to be λ. Thus we get an equation

$$x = x - \frac{f(x)}{M_1}, \tag{12}$$

and the corresponding iterative process has the form

$$x_{n+1} = x_n - \frac{f(x_n)}{M_1}. \tag{13}$$

Reasoning by analogy, we can show that in the case $f'(x) < 0$ and $0 < m_1 \leqslant |f'(x)| \leqslant M_1$ we get an equation

$$x = x + \frac{f(x)}{M_1} \tag{12'}$$

and the corresponding iterative process assumes the form

$$x_{n+1} = x_n + \frac{f(x_n)}{M_1}. \tag{13'}$$

We assume now that in addition to condition (10) we have a relation

$$M_1 \leqslant 3m_1. \tag{14}$$

Then we may require that the inequality $|1 - \lambda f'(x)| \leqslant 1/2$ should be satisfied. Solving it, we get the following restrictions for λ:

$$\frac{1}{2m_1} \leqslant \lambda \leqslant \frac{3}{2M_1}. \tag{15}$$

Geometric interpretation. Consider an equation $f(x) = 0$ [$f(x)$ is a continuous function]. We reduce this equation to the form $x = \varphi(x)$ and construct the graphs of the functions $y = x$ and $y = \varphi(x)$. The abscissa of the point of intersection of the graphs of these functions is the true root ξ (Fig. 5.25).

We choose $x_0 \in [a, b]$ and determine $\varphi(x_0)$. We designate the sequence of points lying on the curve $y = \varphi(x)$ as

A_i ($i = 0, 1, 2, \ldots$) and the sequence of points lying on the straight line $y = x$ as B_i ($i = 1, 2, 3, \ldots$). From the point A_0 $(x_0, \varphi(x_0))$ we draw a straight line, parallel to the x-axis, until it cuts the line $y = x$, and get a point B_1 $(x_1, \varphi(x_0))$.

Indeed, $A_0C_0 = \varphi(x_0) = B_1C_1$ since $A_0B_1 \parallel OC_0$, $B_1C_1 \parallel A_0C_0$. But $OC_1 = B_1C_1$ ($\triangle OC_1B_1$ is right-angled and isosceles since the line $y = x$ is the bisector of the coordinate angle). Consequently, $x_1 = \varphi(x_0)$.

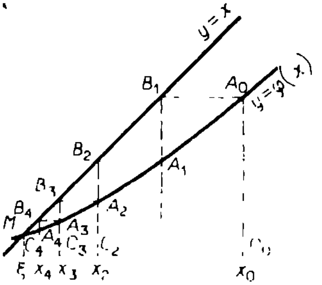


Fig. 5.25

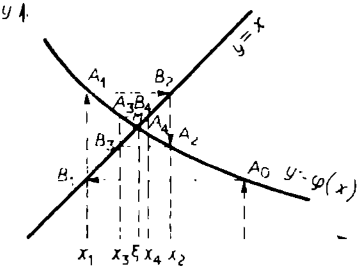


Fig. 5.26

We draw $A_1B_2 \parallel OC_1$ and, repeating the arguments presented above, make sure that $x_2 = \varphi(x_1)$.

Figure 5.25 shows a convergent iterative process. The curve cuts the bisector $y = x$ at the point M with abscissa ξ and, for $x > \xi$, lies under the bisector, and $\varphi'(x)$ satisfies the condition $0 < \varphi'(x) < 1$. The successive approximations $x_0, x_1, \ldots, x_n, \ldots$ (the common abscissas of the points of the graphs of the two functions) *decrease monotonically*. Each successive approximation x_n is closer to the true value of the root than its predecessor x_{n-1}. The polygonal line $A_0B_1A_1B_2A_2 \ldots$ has the form of a staircase.

In Fig. 5.26 the derivative $\varphi'(x) < 0$ but is smaller than unity in absolute value, i.e. $|\varphi'(x)| < 1$. The iterative process converges but the approximations *oscillate* about the exact value of the root. The polygonal line $A_0B_1A_1B_2A_2 \ldots$ has the form of a spiral.

Thus, if in some neighbourhood (a, b) of the root ξ of the equation $x = \varphi(x)$ the derivative $\varphi'(x)$ retains constant sign and the inequality $|\varphi'(x)| \leqslant q < 1$ is satisfied, with $\varphi'(x) > 0$, then the successive approximations $x_n = \varphi(x_{n-1})$ $(n = 1, 2, \ldots)$, $x_0 \in [a, b]$ converge to the root monotonically. When $\varphi'(x) < 0$, the successive approximations oscillate about the root ξ.

Figure 5.27 shows a divergent iterative process. Here $\varphi'(x) > 1$. The curve cuts the bisector $y = x$ at the point M and lies above the bisector for $x > \xi$.

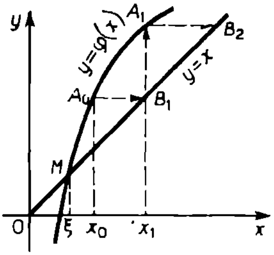


Fig. 5.27

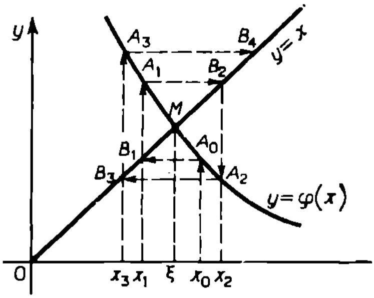


Fig. 5.28

Figure 5.28 illustrates a divergent iterative process for the case $|\varphi'(x)| > 1$. The successive "approximations" recede from the exact value of the root ξ.

Example 1. Use the method of iterations to find the root of the equation $5x^3 - 20x + 3 = 0$, lying on the interval $[0, 1]$ with an accuracy of 10^{-4}.

△ We must reduce the equation to the form $x = \varphi(x)$. There are several ways of doing this, for instance,

$$x = x + (5x^3 - 20x + 3), \quad \text{then} \quad \varphi_1(x) = 5x^3 - 19x + 3,$$

$$x = \sqrt[3]{(20x - 3)/5}, \quad \text{then} \quad \varphi_2(x) = \sqrt[3]{(20x - 3)/5},$$

$$x = (5x^3 + 3)/20, \quad \text{then} \quad \varphi_3(x) = (5x^3 + 3)/20.$$

Let us find out which of the functions obtained must be used to calculate the successive approximations. Recall that if $\varphi(x)$ satisfies the condition $|\varphi'(x)| \leqslant q < 1$ on the interval $[a, b]$, then the iterative process converges. We find that

$$|\varphi_1'(x)| = |15x^2 - 19| > 1 \quad \text{on} \quad [0, 1],$$

$$|\varphi_3'(x)| = 15x^2/20 = 3x^2/4 < 1 \quad \text{on} \quad [0, 1].$$

Consequently, we can use the function $\varphi_3(x)$ and the iterative method to seek the successive approximations from the formula $x_n = (5x_{n-1}^3 + 3)/20$. We take max $\varphi'(x)$ on $[0, 1]$, i.e. $x_0 = 0.75$, as the initial approximation. We employ formula (7) to find the difference between two successive approximations necessary for the specified accuracy to be achieved:

$$|x_n - x_{n-1}| \leqslant \frac{0.0001\cdot(1-0.75)}{0.75} = \frac{0.0001\cdot 0.25}{0.75} = 0.00003.$$

Thus, when the absolute value of the difference $|x_n - x_{n-1}|$ does not exceed 0.00003, the iterative process must be terminated and the specified accuracy assumed to be achieved.

It is convenient to use the following table to carry out the calculations:

Table 5.11

n	x_n	x_n^3	$\varphi(x_n) = x_{n+1}$
0	0.75	0.42188	0.25547
1	0.2555	0.016777	0.154144
2	0.1541	0.005652	0.151413
3	0.1514	0.005443	0.151361
4	0.15136	0.005442	0.151361

At this stage the iterative process may be terminated and $\xi = 0.1514$ may be assumed to be the needed accuracy. ▲

Example 2. Calculate the root of the equation $e^x - x^2 = 0$ with an accuracy of $\varepsilon = 10^{-5}$.

△ We rewrite the equation as $e^x = x^2$ and separate the roots by graphical means. We construct the graphs of the functions

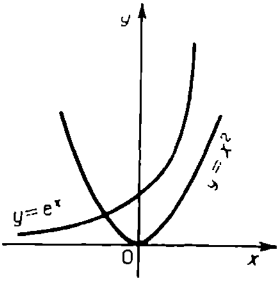


Fig. 5.29

$y = e^x$ and $y = x^2$ (Fig. 5.29). It can be seen from the drawing that the equation $e^x - x^2 = 0$ has one real root which lies on the interval $[-0.8, -0.7]$.

Let us verify whether it is really so We find $f(-0.8)$ and $f(-0.7)$ and have $f(-0.8) = 0.44933 - 0.64 = -0.19067 < 0$,

$f(-0.7) = 0.49659 - 0.49 = 0.00659 > 0$. Since the signs of the function $f(x) = e^x - x^2$ are different at the endpoints of the interval $[-0.8, -0.7]$, the root of the equation is within this interval.

Let us make the interval narrower employing the trial and error method. We find that $f(-0.75) = 0.49237 - 0.56250 < 0$ and $f(-0.7) > 0$. This means that the root is on the interval $[-0.75, 0.7]$. We make the interval narrower still. We have $f(-0.725) = 0.48432 - 0.52562 = -0.4130 < 0$ and $f(-0.7) > 0$. Consequently, the root is on the interval $[-0.725, -0.7]$.

From the equation $e^x = x^2$ we find that $x = -\sqrt{e^x}$ (we take the minus sign before the radical since we know that the root is negative). We rewrite the equation as $x = e^{x/2}$ and find out whether the iterative process is convergent or divergent, i.e. whether the inequality $|\varphi'(x)| < 1$ is satisfied. In this example

$$\varphi(x) = -e^{x/2},\quad \varphi'(x) = (1/2)\, e^{x/2},\quad |\varphi'(-0.725)| = 0.34727,$$

$$|\varphi'(x)| = |\varphi'(-0.7)| = 0.35230.$$

Since $|\varphi'(x)| < 1$, the iterative process converges. We take the number q in formula (7) equal to 0.36. Since $\varepsilon = 10^{-5}$, it follows that

$$|x_n - x_{n-1}| \leqslant \frac{0.00001\,(1-0.36)}{0.36} = 0.000018.$$

Thus the required accuracy will be achieved when the inequality $|x_n - x_{n-1}| \leqslant 0.00002$ is satisfied. We can take any one of the endpoints of the interval $[-0.725, -0.7]$ and any point within it as the zero approximation. We assume that $x_0 = -0.7$.

The calculations can be reduced to the following table:

Table 5.12

n	x_n	$x_n/2$	$e^{x_n/2}$
0	−0.7	−0.35	−0.70460
1	−0.70460	−0.35230	−0.70307
2	−0.70307	−0.35154	−0.70360
3	−0.70360	−0.35180	−0.70342
4	−0.70342	−0.35171	−0.70348
5	−0.70348	−0.35174	−0.70346
6	−0.70346		

Since $|x_6 - x_5| = |-0.70348 - (-0.70346)| = 0.00002$, the required accuracy of the calculations has been achieved and $\xi \cong -0.70346$. ▲

Example 3. Employ the iterative method to calculate the root of the equation $x^3 + 3x^2 - 3 = 0$, lying on the interval $[-2.75, -2.5]$, with an accuracy of 0.001 (see Example 1 in 5.4 and Example 1 in 5.5).

△ We find that $f'(x) = 3x^2 + 6x$. Consequently,

$$M_1 = \max_{[-2.75,\ -2.5]} |f'(x)| = 6.189, \quad m_1 = \min_{[-2.75,\ -2.5]} |f'(x)| = 3.75,$$

$$q = 1 - \frac{m_1}{M_1} = 1 - \frac{3.75}{6.189} < \frac{1}{2}.$$

Since $q < 1/2$, we can use formula (9) to evaluate the error. We assume that $M_1 = 6$ and then $\lambda = 1/6$ and

$$\varphi(x) = x - \lambda f(x) = x - \frac{1}{6}(x^3 + 3x^2 - 3).$$

The corresponding iterative process has the form

$$x_{n+1} = x_n - \frac{1}{6}(x_n^3 + 3x_n^2 - 3),$$

then $|x_{n+1} - x_n| = \frac{1}{6}(x_n^3 + 3x_n^2 - 3)$. The calculations should be terminated as soon as $|x_{n+1} - x_n| < \varepsilon$.

We reduce the calculations to the following table:

Table 5.13

n	x_n	x_n^3	$3x_n^2$	$\frac{1}{6}(x^3 + 3x^2 - 3)$
0	-2.5	-15.625	18.75	0.02
1	-2.52	-16.0030	19.0512	0.0080
2	-2.5280	-16.1559	19.1724	0.0028
3	-2.5308	-16.2096	19.2148	0.0008
4	-2.5316			

Thus we can assume $\xi = -2.532$ to be the approximate value of the root with an accuracy of 0.001. ▲

5.8. General Properties of Algebraic Equations. Determining the Number of Real Roots of an Algebraic Equation

General properties of algebraic equations. We write an nth-degree algebraic equation

$$P_n(x) = a_0 x^n + a_1 x^{n-1} + a_2 x^{n-2} + \dots + a_{n-1} x + a_n = 0, \qquad (1)$$

where $P_n(x)$ is an nth-degree polynomial, n is the highest degree of the unknown, and $a_0, a_1, \dots, a_n$ are real coefficients.

We know that every number ξ which turns the polynomial into zero, i.e. such that $P_n(\xi) = 0$, is a root of the polynomial.

The number ξ is a root of the polynomial $P_n(x)$ if and only if $P_n(x)$ is exactly divisible by $x - \xi$. Recall that if $P_n(x)$ is exactly divisible by $(x - \xi)^k$ $(k \geqslant 1)$, but is not divisible by $(x - \xi^{k+1})$, then ξ is a k-fold root (or a root of multiplicity k) of the polynomial $P_n(x)$. Roots of multiplicity $k = 1$ are simple, or single, roots of a polynomial.

The following theorem, which we give without proof, answers the question whether every polynomial has roots.

Theorem 1 (the fundamental theorem of algebra). *Every polynomial with any numerical coefficients whose degree is not lower than unity has at least one root, which is complex in the general case.*

There is an important corollary of this theorem: *every polynomial* $P_n(x)$ *of degree* n $(n \geqslant 1)$ *with any numerical coefficients has exactly* n *roots, real or complex, if every root is reckoned as many times as is its multiplicity.*

Thus the roots of the algebraic equation (1) may be real as well as complex.

The complex roots of equation (1) possess the property of being *pairwise conjugate*, i.e. if equation (1) has a complex root $\xi = \alpha + \beta i$ (where α and β are real numbers) of multiplicity k, then it also has a complex root $\bar{\xi} = \alpha - \beta i$ also of multiplicity k. These roots are of the same absolute value: $|\xi| = |\bar{\xi}| = \sqrt{\alpha^2 + \beta^2}$.

If equation (1) has complex roots, then their number is even. Therefore every algebraic equation of odd degree with real coefficients has at least one real root.

Before calculating the roots of an algebraic equation, we must: (a) determine the number of roots that equation has, and (b) find the domain of existence of the roots (establish the upper and the lower bound to the roots of the equation). Then we can proceed with determining the roots and making them accurate to a certain value.

Determining the number of real roots of an algebraic equation. We can find out how many positive real roots the algebraic equation (1)

$$P_n(x) = a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \dots + a_{n-1}x + a_n = 0$$

possesses approximately, applying **Descartes' rule of signs**: *the number of positive real roots of the algebraic equation* $P_n(x) = 0$ *with real coefficients (each of which is reckoned according to its degree of multiplicity) either is equal to the number of sign changes in the sequence of the coefficients of the equation* $P_n(x) = 0$ *or is less than the number of sign changes by an even integer (the coefficients equal to zero are not reckoned).*

The number of negative roots of the equation is equal to the number of sign changes in the sequence of coefficients of $P_n(-x)$ or is smaller by an even integer.

If an equation is complete, then the number of its positive roots is equal to the number of variations of sign in the sequence of coefficients or is smaller by an even integer and the number of negative roots is equal to the number of constancies of sign or is smaller by an even integer.

Example 1. Find the number of positive and negative roots of the equation $x^5 - 17x^4 + 12x^3 + 7x^2 - x + 1 = 0$.

△ According to the fundamental theorem of algebra, this equation has five roots (at least one of which is real).

The equation is complete, the sequence of signs of the coefficients being +, —, +, +, —, +. There are four sign changes and this means that there are either four or two positive roots or there are none.

The number of sign constancies is 1, and, consequently, the equation has one negative root. ▲

Example 2. Find the number of positive and negative real roots of the equation $x^6 - 3x^4 + x^3 + x^2 - 1 = 0$.

△ This equation has six roots; the sequence of signs is , —, +, —. There are three sign changes, and, consequently, there are either three positive roots or there is one root. Furthermore, for the polynomial

$$P_n(-x) = x^6 - 3x^4 - x^3 + x^2 - 1$$

the sequence of signs is +, —, —, +, —. We also have three changes of sign here and, therefore, there are either three negative roots or one. ▲

Sturm's theorem allows us to be more precise in determining the number of roots of an algebraic equation.

Since we can always separate the multiple roots of an equation and the common roots of the equations $P_n(x) = 0$ and $P'_n(x) = 0$, we can assume without loosing generality that the equation $P_n(x) = 0$ has only simple roots.

Suppose we have established in some way that all the real roots of the algebraic equation $P_n(x) = 0$ are in the interval (a, b) (a and b are real numbers and are not roots of the equation, $a < b$). We find the first derivative $P'_n(x)$ and divide the polynomial $P_n(x)$ by it. We take the remainder of the division of $P_n(x)$ by $P'_n(x)$ with the opposite sign and denote it by $R_1(x)$.

Then we similarly divide $P'_n(x)$ by $R_1(x)$, take the remainder obtained with the opposite sign and denote it by $R_2(x)$. Dividing $R_1(x)$ by $R_2(x)$ and again taking the remainder with the opposite sign, we get $R_3(x)$. We continue the process of division until we get a remainder which is a constant quantity. We take that quantity also with the opposite sign.

The result is a sequence of functions

$$P_n(x),\ P'_n(x),\ R_1(x),\ R_2(x),\ \ldots,\ R_{m-1}(x),\qquad R_m = \text{const},$$

which is known as *Sturm's system*. We substitute first a and then b for x in this sequence and count the number of sign changes in both cases [we designate the numbers obtained as $W(a)$ and $W(b)$ respectively].

Theorem 2 (Sturm's theorem). *If the real numbers a and b ($a < b$) are not roots of the polynomial $P_n(x)$, which does not have multiple roots, then $W(a) \geqslant W(b)$ and the difference $W(a) - W(b)$ is equal to the number of real roots of the polynomial $P_n(x)$ which lie between a and b.*

Sturm's theorem can be utilized to find the number of negative roots of the equation $P_n(x) = 0$ [i.e. the number of real roots of the equation $P_n(x) = 0$ in the interval $(-\infty, 0)$] or the number of positive roots [in the interval $(0, +\infty)$]. Sturm's theorem is also used to separate roots. The functions entering into Sturm's system can be multiplied and divided by arbitrary positive numbers. This simplifies the calculations considerably when division with a remainder is carried out.

Example 3. Find the number of real roots of the equation $5x^3 - 20x + 3 = 0$, and also separate those roots utilizing Sturm's theorem.

△ We set up a system of Sturm's functions. We have $P_n(x) = 5x^3 - 20x + 3$, $P'_n(x) = 15x^2 - 20$. To determine $R_1(x)$, we multiply $P_n(x)$ by 3 and then divide it by $P_n'(x)$:

$$
\begin{array}{r|l}
 & \quad\quad\quad\quad x \\
\cline{2-2}
15x^2-20 & 15x^3-60x+9 \\
 & \mp 15x^3 \pm 20x \\
\cline{2-2}
 & \quad\quad -40x+9
\end{array}
$$

Hence $R_1(x) = 40x - 9$ (the remainder is taken with the opposite sign). We multiply $P_n'(x)$ by 8 and divide the product by $R_1(x)$:

$$
\begin{array}{r|l}
 & \quad 3x+27 \\
\cline{2-2}
40x-9 & 120x^2-160 \\
 & \mp 120x^2 \pm 27x \\
\cline{2-2}
 & \quad 40\,(27x-160) \\
\hline
 & \quad 40\cdot 27x - 40\cdot 160 \\
\hline
 & \mp 40\cdot 27x \pm 9\cdot 27 \\
\hline
 & \quad\quad\quad -
\end{array}
$$

Since the last remainder is a constant quantity with the minus sign (and in this case we are interested particularly in the sign of the remainder), we change it to the opposite, i.e. to the plus sign.

We compile the following table of the signs of the functions which enter into Sturm's system:

x	$P_n(x)$	$P_n'(x)$	$R_1(x)$	R_2	$W(x)$
$-\infty$	$-$	$+$	$-$	$+$	3
0	$+$	$-$	$-$	$+$	2
$+\infty$	$+$	$+$	$+$	$+$	0

We can see from the table that there are three real roots in the interval $(-\infty, +\infty)$ [since $W(-\infty) - W(+\infty) = 3 - 0 = 3$]. One of them is negative [$W(-\infty) - W(0) = 3 - 2 = 1$] and two are positive [$W(0) - W(+\infty) = 2 - 0 = 2$].

Utilizing Sturm's theorem, we separate the roots diminishing the intervals to the length equal to unity:

x	$P_n(x)$	$P'_n(x)$	$R_1(x)$	$R_2(x)$	$W(x)$
$-\infty$	$-$	$+$	$-$	$+$	3
-3	$-$	$+$	$-$	$+$	3
-2	$+$	$+$	$-$	$+$	2
-1	$+$	$-$	$+$	$+$	2
0	$+$	$-$	$-$	$+$	2
1	$-$	$-$	$+$	$+$	1
2	$+$	$+$	$+$	$+$	0

We can see from this table that the roots lie in the intervals $(-3, -2)$, $(0, 1)$ and $(1, 2)$. ▲

5.9. Finding the Domains of Existence of the Roots of an Algebraic Equation

Rule of annulus. *Assume that we have an algebraic equation*

$$P_n(x) = a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \dots + a_{n-1}x + a_n = 0,$$

where $a_0, a_1, \dots, a_n$ *are real coefficients, and let* $A = \max\{|a_1|, |a_2|, \dots, |a_n|\}$, $B = \max\{|a_0|, |a_1|, \dots, |a_{n-1}|\}$. *Then the roots of the equation are in the annulus* $r < |x| < R$, *where*

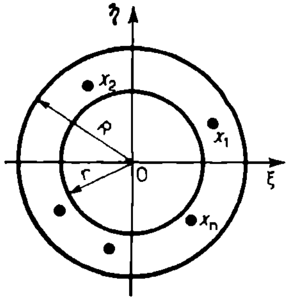


Fig. 5.30

$$r = \frac{1}{1 + B/|a_n|}; \quad R = 1 + \frac{A}{|a_0|}.$$

Here r is the lower bound and R is the upper bound of the positive roots of the algebraic equation $P_n(x) = 0$ and

$-R$, $-r$ are the lower and the upper bound of the negative roots respectively (Fig. 5.30).

Example 1. Determine the bounds of the roots of the equation $5x^3 - 20x + 3 = 0$.

△ Here $|a_0| = 5$, $A = 20$, $|a_n| = 3$, $B = 20$, i.e.

$$R = 1 + \frac{A}{|a_0|} = 1 + \frac{20}{5} = 5;$$

$$r = \frac{1}{1 + B/|a_n|} = \frac{1}{1 + 20/3} = \frac{3}{23} \cong 0.013.$$

Then, if the real roots of the equation $5x^3 - 20x + 3 = 0$ exist (and they are sure to exist since the equation is of an odd degree), they lie in the interval $(-5, 5)$, the negative roots lying in the interval $(-5, -0.013)$ and the positive roots in the interval $(0.013, 5)$. ▲

When solving equations, it is convenient first to establish the bounds of the roots and then use Sturm's theorem. The rule of annulus makes it possible to find the approximate bounds of the roots.

The technique given below allows a more precise estimation of the bounds of the real roots of the algebraic equation $P_n(x) = 0$.

If R_1 is the upper bound of the positive roots of $P_n(x)$, R_2 is the upper bound of the positive roots of $P_n(-x)$, $R_3 > 0$ is the upper bound of the positive roots of $x^n P_n(1/x)$ and R_4 is the upper bound of the positive roots of $x^n P_n(-1/x)$, then all the nonzero real roots of the equation $P_n(x) = 0$ (if they exist) lie within the intervals $(-R_2, -1/R_4)$ and $(1/R_3, R_1)$.

To find the upper bound of the positive roots of an algebraic equation, we can make use of Lagrange's or Newton's method.

Lagrange's method. *If the coefficients of the polynomial*

$$P_n(x) = a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \ldots + a_n$$

satisfy the conditions $a_0 > 0$, a_1, a_2, $\ldots$, $a_{m-1} \geqslant 0$, $a_m < 0$, *then the upper bound of the positive roots of the equation* $P_n(x) = 0$ *can be found from the formula* $R = 1 + \sqrt[m]{B/a_0}$, *where* B *is the greatest of the absolute values of the negative coefficients of* $P_n(x)$.

Example 2. Use Lagrange's method to determine the bounds of the positive and negative roots of the equation $8x^4 - 8x^2 - 32x + 1 = 0$.

△ Here $a_0 = 8 > 0$, $a_1 = 0$, $a_2 = -8 < 0$, $a_3 = -32$, $a_4 = 1$, $m = 2$ (the number of the first negative coefficient), $B = 32$. Consequently, $R_1 = 1 + \sqrt{32/8} = 3$.

Let us consider the polynomial

$$P_n(-x) = 8x^4 - 8x^2 + 32x + 1.$$

We find by analogy that the upper bound of its positive roots is $R_2 = 1 + \sqrt{8/8} = 2$.

Furthermore, for the polynomial

$$x^4 P_n(1/x) = x^4 - 32x^3 - 8x^2 + 8$$

we have $a_0 = 1 > 0$, $a_1 = -32 < 0$, i.e. $m = 1$, $B = 32$, $R_3 = 1 + 32 = 33$.

And for the polynomial

$$x^4 P_n(-1/x) = x^4 + 32x^3 - 8x^2 + 8$$

we have $a_0 = 1 > 0$, $a_1 = 32$, $a_2 = -8$, $a_3 = 0$, $a_4 = 8$, i.e. $m = 2$. Therefore, $R_4 = 1 + \sqrt{8} = 1 + 2\sqrt{2} = 3.828$.

Consequently, if the equation $8x^4 - 8x^2 - 32x + 1 = 0$ has real roots, they are sure to lie in the intervals $(-2, -1/3.828)$ and $(1/33, 3)$. ▲

Newton's method. *If for $x = c$ the polynomial*

$$P_n(x) = a_0x^n + a_1x^{n-1} + \ldots + a_n$$

and its derivatives $P'_n(x)$ and $P''_n(x)$, . . . assume positive values, then c is the upper bound of the positive roots of the equation $P_n(x) = 0$.

Example 3. Use Newton's method to determine the upper bound of the positive roots of the equation $8x^4 - 8x^2 - 32x + 1 = 0$.

△ We find that

$$P(x) = 8x^4 - 8x^2 - 32x + 1,\ P'(x) = 32x^3 - 16x - 32,$$

$$P''(x) = 96x^2 - 16,\ P'''(x) = 192x,\ P^{IV}(x) = 192.$$

We must verify the values of $x > 0$. For $x = c = 1$ we have $P(1) < 0$. This means that we may not continue the verification for $x = 1$ Let us verify the value $x = c = 2$: $P(2) > 0$, $P'(2) > 0$, $P''(2) > 0$, $P'''(2) > 0$, $P^{IV}(2) > 0$. Thus the number 2, i.e. $R = 2$, is the upper bound of the positive roots. We can take the inverse of the number R, i.e. $r = 1/2$, as the lower bound. ▲

5.10. Horner's Method of Approximating Real Roots of an Algebraic Equation

Consider an algebraic equation

$$P(x) = x^n + a_1x^{n-1} + a_2x^{n-2} + \ldots + a_n = 0, \qquad (1)$$

where $a_1, a_2, \ldots, a_n$ are real coefficients of the polynomial. We have to find the real roots of this equation.

We represent the required root of the equation, written in the decimal notation, in the form

$$X = c_0\cdot 10^m + c_1\cdot 10^{m-1} + c_2\cdot 10^{m-2} + \ldots + c_h\cdot 10^{m-k} + \ldots$$

Horner's method consists in the successive determination of the digits of the root c_0, c_1, ... by means of special transformations of equation (1).

If we employ the substitution $x = 10^m\xi$ (for $c_0 > 0$) or $x = -10^m\xi$ (for $c_0 < 0$), then we reduce equation (1) to an equation

$$f_1(\xi) = \xi^n + a'_1\xi^{n-1} + a'_2\xi^{n-2} + \ldots + a'_n = 0,$$

whose root is in the interval (0, 10). Therefore, in what follows we shall consider this particular case. This simplifies the process of computation of the root although it is not obligatory for the employment of Horner's method.

Since $0 < X < 10$, the root of equation (1) can be written in the form

$$X = c_0 + \frac{c_1}{10} + \frac{c_2}{10^2} + \frac{c_3}{10^3} + \ldots = c_0c_1c_2c_3\ldots$$

It is easy to find the first digit c_0 of the root either by the table method or by the use of the interval $[a, b]$ of separation of the root.

Next, applying the transformation

$$x - c_0 = y \qquad (2)$$

to equation (1), we find that $x = y + c_0$, whence it follows that

$$P(x) = P(y + c_0) = -\varphi(y) = 0. \qquad (3)$$

The number $y = 0, c_1c_2c_3 \ldots$ is evidently the root of the last equation. Applying the substitution

$$y = Y/10 \tag{4}$$

to equation (3), we arrive at an equation

$$P_1(Y) = 0, \tag{3'}$$

whose root is $Y = 10y = c_1, c_2c_3 \ldots$. Equation (3′) yields the digit c_1 which is the second digit of the root of equation (1). Then we can again apply substitutions of types (2) and (4), but this time to equation (3′). As a result we get an equation

$$P_2(Z) = 0, \tag{5}$$

whose root is the number $z = c_2, c_3 \ldots$. Then we find the digit c_2.

This procedure can be repeated until we get the required number of digits. Horner's method can be used in combination with some other method, say, the chord method, Newton's method, or the combination of these two methods. We first find several digits of the root by Horner's method and then use other methods to make a closer approximation.

When we have to obtain a small number of the digits of the root, we can use the following technique.

If one of the roots of the equation is considerably smaller than the other roots, then it can be approximately calculated by means of the division of the constant term, taken with the minus sign, by the coefficient of the first power of x.

How can substitutions (2) and (4) be carried out in practice?

To make substitution (4), we must multiply the coefficients $a_1, a_2, \ldots, a_n$ of the initial equation by $10, 10^2, 10^3, \ldots, 10^n$ respectively.

Substitution (2) must be preceded by the expansion of the polynomial $P(x)$ in the powers of $x - c_0$, for which purpose we can use either Taylor's formula or the following relations:

$$
\begin{aligned}
P(x) &= q_1(x)(x - c_0) + r_0, \\
q_1(x) &= q_2(x)(x - c_0) + r_1, \\
q_2(x) &= q_3(x)(x - c_0) + r_2, \\
&\cdots\cdots\cdots\cdots \\
q_{n-2}(x) &= q_{n-1}(x)(x - c_0) + r_{n-2}, \\
q_{n-1}(x) &= (x - c_0) + r_{n-1},
\end{aligned}
\tag{6}
$$

where $q_i(x)$ and r_{i-1} are, respectively, the quotient and the remainder of the division of $q_{i-1}(x)$ by $(x - c_0)$. Successively eliminating $q_i(x)$ from relations (6), we arrive at an identity

$$
\begin{aligned}
P(x) = (x - c_0)^n &+ r_{n-1}(x - c_0)^{n-1} \\
&+ r_{n-2}(x - c_0)^{n-2} + \ldots + r_1(x - c_0) + r_0,
\end{aligned}
\tag{7}
$$

i.e. obtain an expansion of $f(x)$ in the powers of $x - c_0$. Horner's scheme can be used to calculate each of the remainders r_i $(i = 0, 1, \ldots, n - 1)$, i.e. the coefficients of the expansion.

The system of equalities (6) corresponds to the table

c_0

1	a_1	a_2	a_3	...	a_{n-1}	a_n
1	b_1	b_2	b_3	...	b_{n-1}	r_0
1	e_1	e_2	e_3	...	r_1	
.	..	..	..	..	..	
1	r_{n-1}					

where $b_1 = 1 \cdot c_0$, $b_2 = b_1c_0$, $b_3 = b_2c_0$, $\ldots$, $e_1 = 1 \cdot$ $e_2 = e_1c_0$, $e_3 = e_2c_0$, ... If the coefficient of x^n is a_0 then we must write a_0 rather than unity in the first column.

Example. Use Horner's method to find the smallest root of the equation $x^3 + 3x^2 - 3 = 0$ with six significant digits. The roots of the equation have been separated and the smallest of them is on the interval $[-3, -2]$.

$\triangle$ (1) Since the root is negative, we transform the original equation by means of the substitution $x = -\tilde{x}$:

$$-\tilde{x}^3 + 3\tilde{x}^2 - 3 = 0, \quad \text{or} \quad \tilde{x}^3 - 3\tilde{x}^2 + 3 = 0.$$

The required root of the equation $\tilde{x} \in [2, 3]$. Consequently, the first digit of the transformed equation is 2. Then substitutions (2) and (4) have the form $\tilde{x} - 2 = y$ and $y = Y/10$. We use Horner's scheme to make the first substitution:

$c_0 = 2$

1	-3	0	3
1	-1	-2	$\boxed{-1}$
1	1	$\boxed{0}$	
1	$\boxed{3}$		

According to formula (7), we find from this table that $\varphi(y) - y^3 + 3y - 1$ and $P_1(Y) = Y^3 + 30Y^2 - 1000 = 0$.

In what follows, we can use the table to write the two transformations together, retaining the designation of y, i.e. instead of the last relation we shall write $P_1(y) = \tilde{y}^3 + 30y^2 - 1000 = 0$.

We shall find the values of the polynomial $P_1(y)$ for certain values of $y \in [0, 10]$ using Horner's scheme:

y	1	30	0	-1000
5	1	35	175	-125
6	1	36	216	296

Since $P_1(5) < 0$ and $P_1(6) > 0$, it follows that $y \in [5, 6]$. Hence $c_1 = 5$.

(2) To find the next digit, we set up an equation whose coefficients are obtained from Horner's scheme, employing substitution (2) and (4) for $c_1 = 5$:

$c_1 = 5$

1	30	0	-1000
1	35	175	$\boxed{-125}$
1	40	$\boxed{375}$	
1	$\boxed{45}$		

Thus $P_2(y) = y^3 + 450y^2 + 37\,500y - 125\,000 = 0$.

We seek the values of $P_2(y)$ for certain values of $y \in [0, 10]$ using Horner's scheme:

y	1	450	37 500	$-125\,000$
3	1	453	38 859	$-8\,423$
4	1	454	39 316	32 264

Since $P_2(3) < 0$ and $P_2(4) > 0$, it follows that $y \in [3, 4]$. Hence $c_2 = 3$.

(3) We derive an equation for determining the next digits whose coefficients can be found from Horner's scheme:

$c_2 = 3$

1	450	37 500	$-125\,000$
1	453	38 859	$\boxed{-8\,423}$
1	456	$\boxed{40\,227}$	
1	$\boxed{459}$		

Hence $P_3(y) = y^3 + 4590y^2 + 4\,022\,700y - 8\,423\,000 = 0$.

We can apply a special technique described above to the equation obtained. As a result of a three-fold substitution (4), all the roots, except for the required one, have increased approximately 10^3 times. Then the root of the last equation is approximately equal to $8\,423\,000/4\,022\,700 \approx 2.09$. This means that the digits 2, 0 and 9 are the next decimal digits of the root of the original equation. As a result we find the root of the given equation, it is $X = -2.53209$. ▲

Exercises

1. Use analytical means to separate the roots and calculate them with an accuracy of 0.001. Employ the trial and error method.
(a) $x^3 - x + 1 = 0$, (b) $x^3 + 2x - 4 = 0$, (c) $x^4 + 5x - 3 = 0$, (d) $2.2x - 2^x = 0$, (e) $2^x - 2x^2 - 1 = 0$, (f) $2^x - 4x = 0$.

2. Use graphical means to separate the roots and calculate them with an accuracy of 0.001. Employ the chord method:
(a) $x^3 + x - 3 = 0$, (b) $x^3 + 8x - 6 = 0$, (c) $x^3 + 10x - 9 = 0$,
(d) $x^2 - \cos \pi x = 0$, (e) $x^2 - \sin \pi x = 0$, (f) $\log x - \frac{1}{x^2} = 0$.

3. Use Newton's method to find the roots of the following equations with an accuracy of 0.001:
(a) $x^3 - 6x^2 + 9x - 3 = 0$, (b) $x^3 - 12x - 8 = 0$,
(c) $x^3 + 4x^2 - 6 = 0$, (d) $2 \log x - \frac{x}{2} + 1 = 0$,
(e) $x^2 - 20 \sin x = 0$, (f) $x - \cos x = 0$.

4. Use the combination of the chord method and Newton's method to find the roots of the following equations with an accuracy of 0.001:
(a) $x^3+6x-5=0$, (b) $x^3-2x+7=0$,
(c) $x^3-2x^2+x+1=0$,
(d) $1.8x^2-\sin 10x=0$, (e) $\log x-\frac{7}{2x+6}=0$,
(f) $2x\ln x-1=0$.

5. Employing Sturm's theorem, separate the roots of the equations and calculate them with an accuracy of 0.001 using the iterative method:
(a) $x^3+4x-3=0$, (b) $x^4-2x-1=0$,
(c) $x^5-5x+2=0$, (d) $x^4+x-3=0$.

6. Use the iterative method to find the roots of the following equations with an accuracy of 0.001:
(a) $\ln x+(x+1)^3=0$, (b) $\sqrt{x+1}=1/x$, (c) $x-\cos x=0$,
(d) $3x-\cos x-1=0$, (e) $x+\log x=0.5$.

Chapter 6

The Eigenvalues and Eigenvectors of a Matrix

6.1. The Characteristic Polynomial

Consider a square matrix A and a nonzero column vector $\mathbf{x}$:

$$A = \begin{bmatrix} a_{11} & a_{12} \dots a_{1n} \\ a_{21} & a_{22} \dots a_{2n} \\ \dots & \dots \\ a_{n1} & a_{n2} \dots a_{nn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Multiplying the matrix A by the vector $\mathbf{x}$, we obtain a column vector

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

i.e.

$$\mathbf{y} = A\mathbf{x}. \tag{1}$$

If the coordinates y_i ($i = 1, 2, \dots, n$) of the vector $\mathbf{y}$ prove to be proportional to the respective coordinates x_i of the vector $\mathbf{x}$, with the proportionality factor λ, i.e. if $y_i = \lambda x_i$, and, consequently.

$$\mathbf{y} = \lambda\mathbf{x}, \tag{2}$$

then the nonzero column vector $\mathbf{x}$ is an *eigenvector* of the matrix A and the proportionality factor λ is an *eigenvalue* (or *characteristic value*) of the matrix A. Since $\mathbf{y} = A\mathbf{x}$ and $\mathbf{y} = \lambda\mathbf{x}$, it evidently follows that

$$A\mathbf{x} = \lambda\mathbf{x}. \tag{3}$$

Thus, if condition (3) is fulfilled, then the vector $\mathbf{x}$ is the eigenvector of the matrix A corresponding to its eigenvalue λ.

Example 1. Assume that

$$A=\begin{bmatrix}1&2&3\\2&6&-2\\3&4&-1\end{bmatrix},\quad \mathbf{x}=\begin{bmatrix}1\\1\\1\end{bmatrix}.$$

Then

$$A\mathbf{x}=\begin{bmatrix}1&2&3\\2&6&-2\\3&4&-1\end{bmatrix}\begin{bmatrix}1\\1\\1\end{bmatrix}=\begin{bmatrix}6\\6\\6\end{bmatrix}=6\begin{bmatrix}1\\1\\1\end{bmatrix}.$$

Consequently, the number $\lambda = 6$ is an eigenvalue of the matrix A since equality (3) is satisfied:

$$A\mathbf{x}=\lambda\mathbf{x}=6\begin{bmatrix}1\\1\\1\end{bmatrix},$$

and the vector $\mathbf{x}=\begin{bmatrix}1\\1\\1\end{bmatrix}$ is an eigenvector of the matrix A corresponding to the eigenvalue $\lambda=6$.

We rewrite relation (3) in the form $A\mathbf{x}-\lambda\mathbf{x}=0$, or

$$(A-\lambda I)\,\mathbf{x}=0, \tag{4}$$

where I is an identity matrix of the same dimension as the matrix A and 0 is a zero column vector. It is evident that without the factor I in the product λI, equation (4) would be meaningless.

Since

$$\lambda I=\begin{vmatrix}\lambda&0&0&\dots&0\\0&\lambda&0&\dots&0\\\cdot&\cdot&\cdot&\cdot&\cdot\\0&0&0&\dots&\lambda\end{vmatrix},\quad 0=\begin{vmatrix}0\\0\\\vdots\\0\end{vmatrix},$$

we can write equation (4) as

$$\begin{vmatrix}a_{11}-\lambda&a_{12}&\dots&a_{1n}\\a_{21}&a_{22}-\lambda&\dots&a_{2n}\\\cdot&\cdot&\cdot&\cdot\\a_{n1}&a_{n2}&\dots&a_{nn}-\lambda\end{vmatrix}\begin{vmatrix}x_1\\x_2\\\vdots\\x_n\end{vmatrix}=\begin{vmatrix}0\\0\\\vdots\\0\end{vmatrix}. \tag{5}$$

Relation (5) is a homogeneous linear system of equations which has nonzero solutions if and only if its determinant is zero, i.e. when the condition

$$\det (A - \lambda I) = 0 \tag{6}$$

is fulfilled.

Equation (6) is a *characteristic equation* of the matrix A and its left-hand side is a *characteristic polynomial* (or *characteristic determinant*) of the matrix A.

We can also write out the characteristic equation as follows:

$$\begin{vmatrix} a_{11}-\lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22}-\lambda & \dots & a_{2n} \\ \cdot & \cdot & \cdot & \\ a_{n1} & a_{n2} & & a_{nn}-\lambda \end{vmatrix} = 0. \tag{7}$$

If we expand the determinant on the left-hand side of equation (7), we get a polynomial of the nth degree with respect to λ:

$$\begin{aligned} D(\lambda) &= \det (A - \lambda I) \\ &= (-1)^n[\lambda^n - p_1\lambda^{n-1} + p_2\lambda^{n-2} - \dots + (-1)^n p_n]. \end{aligned} \tag{8}$$

The quantity λ, found from equation (8), assumes n values $\lambda_1, \lambda_2, \dots, \lambda_n$, among which equal values may be found. To find the eigenvectors $\mathbf{x}^{(i)}$, which correspond to the eigenvalues λ_i ($i = 1, 2, \dots, n$), we must solve the homogeneous linear system of equations (5) for each value λ_i.

Example 2. Find the eigenvalues and eigenvectors of the matrix $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$.

△ (1) We write the characteristic polynomial of the matrix A and find λ. We have

$$\det (A - \lambda I) = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = \lambda^2 - 4\lambda + 3.$$

The characteristic equation $\lambda^2 - 4\lambda + 3 = 0$ has two roots $\lambda_1 = 1$ and $\lambda_2 = 3$, which are the eigenvalues of the matrix A.

(2) We seek the eigenvector $\mathbf{x}^{(1)} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ corresponding to the value $\lambda_1 = 1$. The matrix equation

$$(A - \lambda_1 I)\,\mathbf{x}^{(1)} = 0, \quad \text{or} \quad \begin{bmatrix} 2-1 & 1 \\ 1 & 2-1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

is equivalent to the system $\begin{cases} x_1+x_2=0 \\ x_1+x_2=0 \end{cases}$, which has an infinite number of solutions of the form $x_1=-x_2$. Setting $x_1=c$ (c is any number), we get $x_2=-c$. Then the required eigenvector can be written as $\mathbf{x}^{(1)}=c\begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

(3) We seek the eigenvector $\mathbf{x}^{(2)}=\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ corresponding to the second eigenvalue $\lambda_2=3$. We have

$$(A-\lambda_2 I)\,\mathbf{x}^{(2)}=0, \quad \text{or} \quad \begin{bmatrix} 2-3 & 1 \\ 1 & 2-3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}=\begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

This matrix equation leads to the system $\begin{cases} -x_1+x_2=0 \\ x_1-x_3=0 \end{cases}$, whence it follows that $x_1=x_2$. Setting $x_1=c$, we find that $x_2=c$. This means that the second eigenvector has the form $\mathbf{x}^{(2)}=c\begin{bmatrix} 1 \\ 1 \end{bmatrix}$. ▲

When eigenvalues and eigenvectors of matrices are sought one of the following two problems must be solved: (1) all the eigenvalues and the corresponding eigenvectors of the matrices must be found, or (2) one or several eigenvalues and the corresponding eigenvectors must be determined.

The first problem consists in expanding the characteristic determinant in an nth-degree polynomial (i.e. in finding the coefficients $p_1, p_2, \ldots, p_n$) and calculating the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ and, finally, finding the coordinates of the eigenvector $\mathbf{x}^{\mathrm{T}}=(x_1, x_2, \ldots, x_n)$.

The second problem consists in finding the eigenvalues λ (one or several of them) using the iterative methods, without expanding the characteristic determinant.

The methods of the first problem are *exact*, i.e. if we apply them to matrices whose elements are defined exactly (by rational numbers) and carry out precise calculations (according to the laws of operations involving common fractions), we shall get the exact values of the coefficients of the characteristic polynomial and the coordinates of the eigenvectors will be expressed by exact formulas in terms of the eigenvalues.

The eigenvectors of a matrix can usually be determined with the use of the intermediate results of the computations carried out to find the coefficients of the characteristic polynomial. It stands to reason that to find an eigenvector corresponding to a certain eigenvalue, this eigenvalue must be already calculated.

The methods of solving the second problem are *iterative*, i.e. the eigenvalues are obtained here as the limit of some number sequences as well as the coordinates of the eigenvectors corresponding to them. Since these methods do not require the calculation of the coefficients of the characteristic polynomial, they are less labour-consuming. In what follows we consider some methods of expanding a characteristic determinant and the iterative methods of finding the eigenvalues of a matrix.

6.2. The Method of Direct Expansion

Let us consider a third-order matrix in order to understand how the coefficients of a characteristic polynomial can be found by a direct expansion of a characteristic determinant. Let

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \quad \det(A - \lambda I) = \begin{vmatrix} a_{11}-\lambda & a_{12} & a_{13} \\ a_{21} & a_{22}-\lambda & a_{23} \\ a_{31} & a_{32} & a_{33}-\lambda \end{vmatrix}$$

We use the rule of triangle to calculate the determinant:

$$\det(A - \lambda I) = \begin{vmatrix} a_{11}-\lambda & a_{12} & a_{13} \\ a_{21} & a_{22}-\lambda & a_{23} \\ a_{31} & a_{32} & a_{33}-\lambda \end{vmatrix}$$

$$= (a_{11}-\lambda)(a_{22}-\lambda)(a_{33}-\lambda) + a_{12}a_{23}a_{31} + a_{13}a_{32}a_{21}$$

$$- a_{13}a_{31}(a_{22}-\lambda) - a_{12}a_{21}(a_{33}-\lambda) - a_{23}a_{32}(a_{11}-\lambda)$$

$$= -\lambda^3 + \lambda^2(a_{11}+a_{22}+a_{33}) - \lambda[(a_{11}a_{22}-a_{12}a_{21})$$

$$+ (a_{22}a_{33}-a_{23}a_{32}) + (a_{11}a_{33}-a_{13}a_{31})] + (a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31}$$

$$+ a_{13}a_{32}a_{21} - a_{13}a_{22}a_{31} - a_{12}a_{21}a_{33} - a_{23}a_{32}a_{11}$$

$$= (-1)^3 \cdot \left[\lambda^3 - \lambda^2(a_{11}+a_{22}+a_{33}) + \lambda\left(\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}\right.\right.$$

$$\left.\left. + \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} + \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix}\right) - \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}\right],$$

or

$$\det(A - \lambda I) = (-1)^3(\lambda^3 - p_1\lambda^2 + p_2\lambda - p_3) = 0.$$

The coefficient p_1 here is the sum of the diagonal elements of the matrix A. It is known as the *trace* of the matrix and is designated as Tr A:

$$p_1 = \operatorname{Tr} A = a_{11} + a_{22} + a_{33};$$

the coefficient p_2 is the sum of all the principal minors of the second order of the matrix A:

$$p_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} + \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} + \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix}$$

(recall that the principal minors of the second, the third, ... the nth order are minors the elements of whose principal diagonals are the elements of the principal diagonal of the determinant det A); the coefficient

$$p_3 = \det A = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}.$$

In general, we have to expand the determinant det $(A - \lambda I)$ in an nth-degree polynomial

$$D(\lambda) = -(1)^n [\lambda^n - p_1\lambda^{n-1} + p_2\lambda^{n-2} - \ldots + (-1)^n p_n]$$

then the coefficients $p_1, p_2, \ldots, p_n$ can be found from the following formulas:

$p_1 = \sum_{i=1}^{n} a_{ii} = \operatorname{Tr} A$ which is the sum of all the diagonal elements of the matrix A,

$p_2 = \sum_{\alpha<\beta} \begin{vmatrix} a_{\alpha\alpha} & a_{\alpha\beta} \\ a_{\beta\alpha} & a_{\beta\beta} \end{vmatrix}$ which is the sum of all the principal minors of the second order of the matrix A,

$p_3 = \sum_{\alpha<\beta<\gamma} \begin{vmatrix} a_{\alpha\alpha} & a_{\alpha\beta} & a_{\alpha\gamma} \\ a_{\beta\alpha} & a_{\beta\beta} & a_{\beta\gamma} \\ a_{\gamma\alpha} & a_{\gamma\beta} & a_{\gamma\gamma} \end{vmatrix}$ which is the sum of all the principal minors of the third order of the matrix A.

$p_n = \det A$ which is the determinant of the matrix A.

The number of principal minors of order k of the matrix A is

$$\binom{n}{k} = \frac{n(n-1)(n-2)\ldots(n-k+1)}{k!} \quad (k = 1, 2, \ldots, n).$$

The method of direct expansion is very labour-consuming and is only used to find the characteristic polynomials for low-order matrices.

Example 1. Use the method of direct expansion to find the characteristic polynomial of the matrix

$$A=\begin{bmatrix} -4 & -3 & 1 & 1 \\ 2 & 0 & 4 & -1 \\ 1 & 1 & 2 & -2 \\ 1 & 1 & -1 & -1 \end{bmatrix}.$$

(1) We find that

$$p_1 = \operatorname{Tr} A = a_{11} + a_{22} + a_{33} + a_{44} = -4 + 0 + 2 - 1 = -3.$$

(2) We have $p_2 = \sum\limits_{\alpha<\beta} \begin{vmatrix} a_{\alpha\alpha} & a_{\alpha\beta} \\ a_{\beta\alpha} & a_{\beta\beta} \end{vmatrix}$. The number of second-order principal minors of a fourth-order matrix is $\binom{4}{2} = \frac{4\cdot 3}{1\cdot 2} = 6$. Writing out all these minors and adding them together, we get

$$p_2 = \underbrace{\begin{vmatrix} -4 & -3 \\ 2 & 0 \end{vmatrix}}_{\alpha=1;\ \beta=2} + \underbrace{\begin{vmatrix} -4 & 1 \\ 1 & 2 \end{vmatrix}}_{\alpha=1;\ \beta=3} + \underbrace{\begin{vmatrix} -4 & 1 \\ 1 & -1 \end{vmatrix}}_{\alpha=1;\ \beta=4} + \underbrace{\begin{vmatrix} 0 & 4 \\ 1 & 2 \end{vmatrix}}_{\alpha=2;\ \beta=3}$$
$$+ \underbrace{\begin{vmatrix} 0 & -1 \\ 1 & -1 \end{vmatrix}}_{\alpha=2,\ \beta=4} + \underbrace{\begin{vmatrix} 2 & -2 \\ -1 & -1 \end{vmatrix}}_{\alpha=3;\ \beta=4} = -7.$$

(3) We have $p_3 = \sum\limits_{\alpha<\beta<\gamma} \begin{vmatrix} a_{\alpha\alpha} & a_{\alpha\beta} & a_{\alpha\gamma} \\ a_{\beta\alpha} & a_{\beta\beta} & a_{\beta\gamma} \\ a_{\gamma\alpha} & a_{\gamma\beta} & a_{\gamma\gamma} \end{vmatrix}$. The number of third-order principal minors of a fourth-order matrix is $\binom{4}{3} = \frac{4\cdot 3\cdot 2}{1\cdot 2\cdot 3} = 4$. Consequently,

$$p_3 = \underbrace{\begin{vmatrix} -4 & -3 & 1 \\ 2 & 0 & 4 \\ 1 & 1 & 2 \end{vmatrix}}_{\alpha=1;\ \beta=2;\ \gamma=3} + \underbrace{\begin{vmatrix} -4 & -3 & 1 \\ 2 & 0 & -1 \\ 1 & 1 & -1 \end{vmatrix}}_{\alpha=1;\ \beta=2;\ \gamma=4} + \underbrace{\begin{vmatrix} -4 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & -1 & -1 \end{vmatrix}}_{\alpha=1;\ \beta=3;\ \gamma=4}$$
$$+ \underbrace{\begin{vmatrix} 0 & 4 & -1 \\ 1 & 2 & -2 \\ 1 & -1 & -1 \end{vmatrix}}_{\alpha=2;\ \beta=3;\ \gamma=4} = 24.$$

(4) And finally we find that

$$p_4=\det A=\begin{vmatrix} -4 & -3 & 1 & 1 \\ 2 & 0 & 4 & -1 \\ 1 & 1 & 2 & -2 \\ 1 & 1 & -1 & -1 \end{vmatrix}=-15.$$

(5) The final result is

$$\begin{aligned} D(\lambda) &= \lambda^4-p_1\lambda^3+p_2\lambda^2-p_3\lambda+p_4 \\ &= \lambda^4+3\lambda^3-7\lambda^2-24\lambda-15. \ \blacktriangle \end{aligned}$$

Example 2. Expand the characteristic polynomial of the matrix

$$A=\begin{bmatrix} 3 & 1 & 0 \\ -4 & -1 & 0 \\ 4 & -8 & -2 \end{bmatrix}$$

and find any of the eigenvalues and the corresponding eigenvector.

△ (1) We have

$$p_1=\operatorname{Tr}A=3-1-2=0;$$

$$p_2=\sum_{\alpha<\beta}\begin{vmatrix} a_{\alpha\alpha} & a_{\alpha\beta} \\ a_{\beta\alpha} & a_{\beta\beta} \end{vmatrix}=\underbrace{\begin{vmatrix} 3 & 1 \\ -4 & -1 \end{vmatrix}}_{\alpha=1;\ \beta=2}+\underbrace{\begin{vmatrix} 3 & 0 \\ 4 & -2 \end{vmatrix}}_{\alpha=1;\ \beta=3}+\underbrace{\begin{vmatrix} -1 & 0 \\ -8 & -2 \end{vmatrix}}_{\alpha=2;\ \beta=3}$$

$$=1-6+2=-3;$$

$$p_3=\det A=\begin{vmatrix} 3 & 1 & 0 \\ -4 & -1 & 0 \\ 4 & -8 & -2 \end{vmatrix}=-2.$$

Consequently, $D(\lambda)=(-1)^3(\lambda^3-3\lambda+2)$, $\lambda^3-3\lambda+2=0$. One of the eigenvalues of the matrix A is $\lambda=1$.

(2) Let us find the eigenvector $\mathbf{x}=\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ corresponding to $\lambda=1$.

The matrix equation

$$(A-\lambda I)\mathbf{x}=0,\quad \text{or}\quad \begin{bmatrix} 3-1 & 1 & 0 \\ -4 & -1-1 & 0 \\ 4 & -8 & -2-1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}=\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

is equivalent to the system

$$\begin{cases} 2x_1+x_2=0, \\ -4x_1-2x_2=0, \\ 4x_1-8x_2-3x_3=0. \end{cases}$$

The first and the second equations are proportional here and, therefore, discarding the second equation, we get a system

$$\begin{cases} 2x_1+x_2 \qquad\quad =0, \\ 4x_1-8x_2-3x_3=0. \end{cases}$$

The minor $d=\begin{vmatrix} 2 & 1 \\ 4 & -8 \end{vmatrix}=-20$ is a base minor, x_1 and x_2 are base unknowns, x_3 is a free unknown. Using then Cramer's rule, we find that

$$x_1=\frac{d_1}{d}=\frac{\begin{vmatrix} 0 & 1 \\ 3x_3 & -8 \end{vmatrix}}{-20}=\frac{3x_3}{20},$$

$$x_2=\frac{d_2}{d}=\frac{\begin{vmatrix} 2 & 0 \\ 4 & 3x_3 \end{vmatrix}}{-20}=\frac{6x_3}{-20}.$$

Setting $x_3=20$, we find that $x_1=3$, $x_2=-6$. Thus $\mathbf{x}^{(1)}=c\begin{bmatrix} 3 \\ -6 \\ 20 \end{bmatrix}$ is the required eigenvector. ▲

6.3. Krylov's Method of Expansion of a Characteristic Determinant

Krylov's method is based on the property of a square matrix to turn its characteristic polynomial into zero.

According to the **Hamilton-Cayley theorem**, *every square matrix is a root of its characteristic polynomial and,* consequently, *turns it into zero.*

Let

$$D(\lambda)=\det(A-\lambda I)$$
$$=(-1)^n(\lambda^n+p_1\lambda^{n-1}+p_2\lambda^{n-2}+\ldots+p_n) \qquad (1)$$

be a characteristic polynomial of the matrix A. Replacing the quantity λ in relation (1) by $A=[a_{ij}]$ (where $i=1, 2, \ldots, n$, $j=1, 2, \ldots, n$), we obtain

$$A^n+p_1A^{n-1}+p_2A^{n-2}+\ldots+p_nI=0. \qquad (2)$$

We take an arbitrary nonzero vector

$$\mathbf{y}^{(0)} = \begin{bmatrix} y_1^{(0)} \\ y^{(0)} \\ \vdots \\ y_n^{(0)} \end{bmatrix} \tag{3}$$

and postmultiply both sides of relation (2) by $\mathbf{y}^{(0)}$:

$$A^n\mathbf{y}^{(0)} + p_1A^{n-1}\mathbf{y}^{(0)} + p_2A^{n-2}\mathbf{y}^{(0)} + \ldots + p_n\mathbf{y}^{(0)} = 0. \tag{4}$$

Now we set

$$A\mathbf{y}^{(k-1)} = \mathbf{y}^{(k)} \; (k = 1, 2, \ldots, n), \tag{5}$$

i.e.

$$\begin{aligned} \mathbf{y}^{(1)} &= A\mathbf{y}^{(0)}, \\ \mathbf{y}^{(2)} &= A\mathbf{y}^{(1)} = A^2\mathbf{y}^{(0)}, \\ &\ldots\ldots\ldots\ldots \\ \mathbf{y}^{(n)} &= A\mathbf{y}^{n-1} = A^n\mathbf{y}^{(0)}. \end{aligned}$$

Then relation (4) assumes the form

$$\mathbf{y}^{(n)} + p_1\mathbf{y}^{(n-1)} + p_2\mathbf{y}^{(n-2)} + \ldots + p_n\mathbf{y}^{(0)} = 0, \tag{6}$$

or

$$p_1\mathbf{y}^{(n-1)} + p_1\mathbf{y}^{(n-2)} + \ldots + p_n\mathbf{y}^{(0)} = -\mathbf{y}^{(n)},$$

or

$$p_1\begin{bmatrix} y_1^{(n-1)} \\ y_2^{(n-1)} \\ \vdots \\ y_n^{(n-1)} \end{bmatrix} + p_2\begin{bmatrix} y_1^{(n-2)} \\ y_2^{(n-2)} \\ \vdots \\ y_n^{(n-2)} \end{bmatrix} +$$

$$\ldots + p_n\begin{bmatrix} y_1^{(0)} \\ y_2^{(0)} \\ \vdots \\ y_n^{(0)} \end{bmatrix} = -\begin{bmatrix} y_1^{(n)} \\ y_2^{(n)} \\ \vdots \\ y_n^{(n)} \end{bmatrix},$$

i.e.

$$\left\{\begin{array}{l} p_1 y_1^{(n-1)} + p_2 y_1^{(n-2)} + \ldots + p_n y_1^{(0)} = -y_1^{(n)}, \\ p_1 y_2^{(n-1)} + p_2 y_2^{(n-2)} + \ldots + p_n y_2^{(0)} = -y_2^{(n)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ p_1 y_n^{(n-1)} + p_2 y_n^{(n-2)} + \ldots + p_n y_n^{(0)} = -y_n^{(n)}, \end{array}\right. \tag{7}$$

or, finally, in matrix form

$$\begin{bmatrix} y_1^{(n-1)} & y_1^{(n-2)} & \ldots & y_1^{(0)} \\ y_2^{(n-1)} & y_2^{(n-2)} & \ldots & y_2^{(0)} \\ \ldots & \ldots & \ldots & \ldots \\ y_n^{(n-1)} & y_n^{(n-2)} & \ldots & y_n^{(0)} \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix} = -\begin{bmatrix} y_1^{(n)} \\ y_2^{(n)} \\ \vdots \\ y_n^{(n)} \end{bmatrix}. \tag{8}$$

The vectors $\mathbf{y}_i^{(1)}, \mathbf{y}_i^{(2)}, \ldots, \mathbf{y}_i^{(n)}$ can be found from the formulas

$$\begin{aligned} \mathbf{y}_i^{(1)} &= \sum_{j=1}^{n} a_{ij}\mathbf{y}_j^{(0)} = A\mathbf{y}^{(0)}, \\ \mathbf{y}_i^{(2)} &= \sum_{j=1}^{n} a_{ij}\mathbf{y}_j^{(1)} = A\mathbf{y}^{(1)}, \\ &\ldots\ldots\ldots\ldots \\ \mathbf{y}_i^{(n)} &= \sum_{j=1}^{n} a_{ij}\mathbf{y}_j^{(n-1)} = A\mathbf{y}^{(n-1)} \ (i = 1, 2, \ldots, n), \end{aligned} \tag{9}$$

the coordinates of the initial vector (3) being arbitrary. If the linear system (7) has a unique solution, then its roots $p_1, p_2, \ldots, p_n$ are the coefficients of the characteristic polynomial (1). We can use Gauss' elimination to find this solution.

Example 1. Use Krylov's method to expand the characteristic determinant of the matrix

$$A = \begin{bmatrix} -4 & -3 & 1 & 1 \\ 2 & 0 & 4 & -1 \\ 1 & 1 & 2 & -2 \\ 1 & 1 & -1 & -1 \end{bmatrix}.$$

$\triangle$ (1) We choose an initial vector $\mathbf{y}^{(0)} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$.

Table 6.1

p_1	p_2	p_3	p_4	Constant terms	Σ_1	Σ_2
$\boxed{-39}$	12	−4	1	−120	−150	
20	−5	2	0	47	64	
11	−2	1	0	23	33	
13	−4	1	0	43	53	
1	−4/13	4/39	−1/39	40/13	50/13	
	$\boxed{15/13}$	−2/39	20/39	−189/13	−168/13	−168/13
	18/13	−5/39	11/39	−141/13	−121/13	−121/13
	0	−1/3	−1/3	3	3	3
	1	−2/45	4/9	−53/5	−56/5	−56/5
		$\boxed{-1/15}$	−1/3	33/5	31/5	31/5
		−1/3	1/3	3	3	3
		1	5	−99	−93	−93
			$\boxed{2}$	−30	−28	−28
			1	−15	−14	−14
			1	$p_4=-15$	$\bar{p}_4=-14$	−14
		1		$p_3=-24$	$\bar{p}_3=-23$	−23
	1			$p_2=-7$	$\bar{p}_2=-6$	−6
1				$p_1=3$	$\bar{p}_1=4$	4

(2) Using formulas (9), we determine the coordinates of the vectors $\mathbf{y}^{(k)} = A\mathbf{y}^{(k-1)}$ ($k = 1, 2, 3, 4$):

$$\mathbf{y}^{(1)} = A\mathbf{y}^{(0)} = \begin{bmatrix} -4 & -3 & 1 & 1 \\ 2 & 0 & 4 & -1 \\ 1 & 1 & 2 & -2 \\ 1 & 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -4 \\ 2 \\ 1 \\ 1 \end{bmatrix},$$

$$\mathbf{y}^{(2)} = A\mathbf{y}^{(1)} = \begin{bmatrix} -4 & -3 & 1 & 1 \\ 2 & 0 & 4 & -1 \\ 1 & 1 & 2 & -2 \\ 1 & 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} -4 \\ 2 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 12 \\ -5 \\ -2 \\ -4 \end{bmatrix},$$

$$\mathbf{y}^{(3)} = A\mathbf{y}^{(2)} = \begin{bmatrix} -4 & -3 & 1 & 1 \\ 2 & 0 & 4 & -1 \\ 1 & 1 & 2 & -2 \\ 1 & 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} 12 \\ -5 \\ -2 \\ -4 \end{bmatrix} = \begin{bmatrix} -39 \\ 20 \\ 11 \\ 13 \end{bmatrix},$$

$$\mathbf{y}^{(4)} = A\mathbf{y}^{(3)} = \begin{bmatrix} -4 & -3 & 1 & 1 \\ 2 & 0 & 4 & -1 \\ 1 & 1 & 2 & -2 \\ 1 & 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} -39 \\ 20 \\ 11 \\ 13 \end{bmatrix} = \begin{bmatrix} 120 \\ -47 \\ -23 \\ -43 \end{bmatrix}.$$

(3) Then we set up a matrix equation

$$\begin{bmatrix} -39 & 12 & -4 & 1 \\ 20 & -5 & 2 & 0 \\ 11 & -2 & 1 & 0 \\ 13 & -4 & 1 & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix} = -\begin{bmatrix} 120 \\ -47 \\ -23 \\ -43 \end{bmatrix}$$

and write a system of form (7):

$$\begin{cases} -39p_1 + 12p_2 - 4p_3 + p_4 = -120, \\ 20p_1 - 5p_2 + 2p_3 = 47, \\ 11p_1 - 2p_2 + p_3 = 23, \\ 13p_1 - 4p_2 + p_3 = 43. \end{cases}$$

We solve this system using Gauss' elimination (see **Table 6.1**).

Thus we have

$$D(\lambda) = \det(A - \lambda I) = \lambda^4 + p_1\lambda^3 + p_2\lambda^2 + p_3\lambda + p_4$$
$$= \lambda^4 + 3\lambda^3 - 7\lambda^2 - 24\lambda - 15. \ \blacktriangle$$

Now if the linear system (7) does not have a unique solution, the initial vector must be altered.

Example 2. Use Krylov's method to expand the characteristic determinant of the matrix

$$A=\begin{bmatrix}1 & -1 & -1 & 2\\ 2 & 3 & 0 & -4\\ 1 & 1 & -2 & -2\\ 1 & 1 & 0 & -1\end{bmatrix}.$$

∧ (1) We take $\mathbf{y}^{(0)}=\begin{bmatrix}1\\0\\0\\0\end{bmatrix}$ as the initial vector and obtain

$$\mathbf{y}^{(1)}=A\mathbf{y}^{(0)}=\begin{bmatrix}1\\2\\1\\1\end{bmatrix};\quad \mathbf{y}^{(2)}=A\mathbf{y}^{(2)}=\begin{bmatrix}0\\4\\-1\\2\end{bmatrix};$$

$$\mathbf{y}^{(3)}=A\mathbf{y}^{(2)}=\begin{bmatrix}1\\4\\2\\2\end{bmatrix};\quad \mathbf{y}^{(4)}=A\mathbf{y}^{(3)}=\begin{bmatrix}-1\\6\\-3\\3\end{bmatrix}.$$

(2) We set up a matrix equation

$$\begin{bmatrix}1 & 0 & 1 & 1\\ 4 & 4 & 2 & 0\\ 2 & -1 & 1 & 0\\ 2 & 2 & 1 & 0\end{bmatrix}\begin{bmatrix}p_1\\p_2\\p_3\\p_4\end{bmatrix}=-\begin{bmatrix}1\\6\\-3\\3\end{bmatrix},$$

whence we obtain a system of equations

$$\begin{cases}p_1+p_3+p_4=1,\\ 4p_1+4p_2+2p_3=-6,\\ 2p_1-p_2+p_3=3,\\ 2p_1+2p_2+p_3=-3.\end{cases}$$

We solve this system using the scheme of unique division (see Table 6.2).

Since the pivot element is zero, it is impossible to continue the calculations using this scheme.

(3) To obtain the unique solution, we change the initial vector. Setting $\mathbf{y}^{(0)}=\begin{bmatrix}0\\0\\0\\1\end{bmatrix}$ we find that

$$\mathbf{y}^{(1)} = A\mathbf{y}^{(0)} = \begin{bmatrix} 2 \\ -4 \\ -2 \\ -1 \end{bmatrix}; \quad \mathbf{y}^{(2)} = A\mathbf{y}^{(1)} = \begin{bmatrix} 6 \\ -4 \\ 4 \\ -1 \end{bmatrix};$$

$$\mathbf{y}^{(3)} = A\mathbf{y}^{(2)} = \begin{bmatrix} 4 \\ 4 \\ -4 \\ 3 \end{bmatrix}; \quad \mathbf{y}^{(4)} = A\mathbf{y}^{(3)} = \begin{bmatrix} 10 \\ 8 \\ 10 \\ 5 \end{bmatrix}.$$

The matrix equation

$$\begin{bmatrix} 4 & 6 & 2 & 0 \\ 4 & -4 & -4 & 0 \\ -4 & 4 & -2 & 0 \\ 3 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix} = -\begin{bmatrix} 10 \\ 8 \\ 10 \\ 5 \end{bmatrix}$$

Table 6.2

p_1	p_2	p_3	p_4	Constant terms	Σ_1	Σ_2
$\boxed{1}$	0	1	1	1	4	
4	4	2	0	−6	4	
2	−1	1	0	3	5	
2	2	1	0	−3	2	
1	0	1	1	1	4	
	$\boxed{4}$	−2	−4	−10	−12	−12
	−1	−1	−2	1	−3	−3
	2	−1	−2	−5	−6	−6
	1	−0.5	−1	−3	−3.5	−3.5
		$\boxed{-1.5}$	−3	−2	−6.5	−6.5
		0	0	1	1	1
		1	2	1.333	4.333	4.333
			$\boxed{0}$	1	1	1

leads to the system

$$\begin{cases} 4p_1+6p_2+2p_3 = -10 \\ 4p_1-4p_2-4p_3 = -8, \\ -4p_1+4p_2-2p_3 = -10 \\ 3p_1-p_2-p_3+p_4 = -5, \end{cases}$$

or

$$\begin{cases} 2p_1+3p_2+p_3 = -5, \\ p_1-p_2-p_3 = -2, \\ -2p_1+2p_2-p_3 = -5, \\ 3p_1-p_2-p_3+p_4 = -5, \end{cases}$$

Table 6.3

p_1	p_2	p_3	p_4	Constant terms	Σ_1	Σ_2
$\boxed{2}$	3	1	0	−5	1	
1	−1	−1	0	2	−3	
−2	2	−1	0	−5	−6	
3	−1	−1	1	−5	−3	
1	1.5	0.5	0	−2.5	0.5	
	$\boxed{-2.5}$	−1.5	0	−0.5	−3.5	−3.5
	5	0	0	−10	−5	−5
	−5.5	−2.5	1	2.5	−4.5	−4.5
	1	0.6	0	−0.2	1.4	1.4
		$\boxed{-3}$	0	−9	−12	−12
		0.8	1	1.4	3.2	3.2
		1	0	3	4	4
			$\boxed{1}$	−1	0	0
			1	$p_4=-1$	$\bar{p}_4=0$	0
		1		$p_3=3$	$\bar{p}_3=4$	4
	1			$p_2=-2$	$\bar{p}_2=-1$	−1
1				$p_1=-1$	$\bar{p}_1=0$	0

which can be solved with the use of the scheme of unique division (see Table 6.3).

Consequently,

$$D(\lambda) = \lambda^4 - \lambda^3 - 2\lambda^2 + 3\lambda - 1. \blacktriangle$$

6.4. Using Krylov's Method for Calculation of Eigenvectors

If the coefficients $p_1, p_2, \ldots, p_n$ and the roots $\lambda_1, \lambda_2, \ldots, \lambda_4$ of the characteristic polynomial are known, then Krylov's method makes it possible to find the corresponding eigenvectors from the following formula:

$$\mathbf{x}^{(i)} = \mathbf{y}^{(n-1)} + q_{1i}\mathbf{y}^{(n-2)} + \ldots + q_{n-1,\,i}\,\mathbf{y}^{(0)} \quad (i = 1, 2, \ldots, n). \qquad (1)$$

Here $\mathbf{y}^{(n-1)}, \mathbf{y}^{(n-2)}, \ldots, \mathbf{y}^{(0)}$ are vectors used for seeking the coefficients $p_1, p_2, \ldots, p_n$ by Krylov's method and the coefficients q_{ji} ($j = 1, 2, \ldots, n-1$, $i = 1, 2, \ldots, n$) can be found using Horner's scheme:

$$q_{0i} = 1, \quad q_{ji} = \lambda_i q_{j-1,\,i} + p_j. \qquad (2)$$

Example. **Use Krylov's method to find the eigenvectors of the matrix**

$$A = \begin{bmatrix} 1 & -1 & -1 & 2 \\ 2 & 3 & 0 & -4 \\ 1 & 1 & -2 & -2 \\ 1 & 1 & 0 & -1 \end{bmatrix}.$$

△The characteristic polynomial of the matrix A is known:

$$\det(A - \lambda I) = \lambda^4 - \lambda^3 - 2\lambda^2 + 3\lambda - 1$$

(see Example 2 in 6.3) and the eigenvalues are $\lambda_1 = \lambda_2 = 1$, $\lambda_3 = 0.618$, $\lambda_4 = -1.618$. To find the eigenvectors, we use for-

Table 6.4

λ_i	$p_0 = 1$	$p_1 = 1$	$p_2 = -2$	$p_3 = 3$
$\lambda_1 = 1$	$q_{01} = 1$	$q_{11} = 0$	$q_{21} = -2$	$q_{31} = 1$
$\lambda_2 = 1$	$q_{02} = 1$	$q_{12} = 0$	$q_{22} = -2$	$q_{32} = 1$
$\lambda_3 = 0.618$	$q_{03} = 1$	$q_{13} = -0.382$	$q_{23} = -2.236$	$q_{33} = 1.618$
$\lambda_4 = -1.618$	$q_{04} = 1$	$q_{14} = -2.618$	$q_{24} = 2.236$	$q_{34} = -0.618$

mula (1):

$$\mathbf{x}^{(i)} = \mathbf{y}^{(3)} + q_{1i}\mathbf{y}^{(2)} + q_{2i}\mathbf{y}^{(1)} + q_{3i}\mathbf{y}^{(0)}.$$

Here $q_{0i} = 1$ and the coefficients q_{ji} ($j = 1, 2, 3$, $i = 1, 2, 3, 4$) can be found with the use of Horner's scheme (see Table 6.4).

We employ the expressions for the vectors $\mathbf{y}^{(0)}$, $\mathbf{y}^{(1)}$, $\mathbf{y}^{(2)}$, $\mathbf{y}^{(3)}$, which we have found in Example 2 in 6.3, and obtain

$$\mathbf{x}^{(1)} = \mathbf{x}^{(2)} = \begin{bmatrix} 4 \\ 4 \\ -4 \\ 3 \end{bmatrix} + 0\begin{bmatrix} 6 \\ -4 \\ 4 \\ -1 \end{bmatrix} - 2\cdot\begin{bmatrix} 2 \\ -4 \\ -2 \\ -1 \end{bmatrix}$$

$$+1\cdot\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 12 \\ 0 \\ 6 \end{bmatrix} \approx \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0.5 \end{bmatrix};$$

$$\mathbf{x}^{(3)} = \begin{bmatrix} 4 \\ 4 \\ -4 \\ 3 \end{bmatrix} - 0.382\begin{bmatrix} 6 \\ -4 \\ 4 \\ -1 \end{bmatrix} - 2.236\begin{bmatrix} 2 \\ -4 \\ -2 \\ -1 \end{bmatrix}$$

$$+1.618\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \approx \begin{bmatrix} -2.764 \\ 14.472 \\ -1.056 \\ 7.236 \end{bmatrix} = \begin{bmatrix} 0.19 \\ 1 \\ -0.07 \\ 0.50 \end{bmatrix};$$

$$\mathbf{x}^{(4)} = \begin{bmatrix} 4 \\ 4 \\ -4 \\ 3 \end{bmatrix} - 2.618\begin{bmatrix} 6 \\ -4 \\ 4 \\ -1 \end{bmatrix} + 2.236\begin{bmatrix} 2 \\ -4 \\ -2 \\ -1 \end{bmatrix}$$

$$-0.618\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} -7.236 \\ 5.528 \\ -18.944 \\ 2.764 \end{bmatrix} \approx \begin{bmatrix} -0.38 \\ 0.29 \\ 1 \\ 0.15 \end{bmatrix}. \blacktriangle$$

6.5. The Leverrier-Faddeev Method

This method was suggested by Leverrier and then simplified by the Soviet mathematician Faddeev. The method of Leverrier is based on Newton's formulas for the sums of the powers of the roots of an algebraic equation and consists in the following. Assume that

$$\det(A - \lambda I) = \lambda^n + p_1\lambda^{n-1} + \ldots + p_n \qquad (1)$$

is a characteristic polynomial of the matrix $A = [a_{ij}]$ (where $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, n$) and $\lambda_1, \lambda_2, \ldots, \lambda_n$ is a complete set of roots of polynomial (1). Consider the sums

$$S_k = \lambda_1^k + \lambda_2^k + \ldots + \lambda_n^k \ (k = 1, 2, \ldots, n),$$

i.e.

$$\begin{aligned} S_1 &= \lambda_1 + \lambda_2 + \ldots + \lambda_n = \mathrm{Tr}\, A, \\ S_2 &= \lambda_1^2 + \lambda_2^2 + \ldots + \lambda_n^2 = \mathrm{Tr}\, A^2, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ S_n &= \lambda_1^n + \lambda_2^n + \ldots + \lambda_n^n = \mathrm{Tr}\, A^n \end{aligned}$$

(each sum S_k is a trace of the matrix A^k). Then, for $k \leqslant n$, there hold *Newton's formulas*

$$S_k + p_1 S_{k-1} + \ldots + p_{k-1} S_1 = -kp_k,$$

whence we find that

$$\begin{aligned} p_1 &= -S_1 && \text{for } k = 1 \\ p_2 &= -\frac{1}{2}(S_2 + p_1 S_1) && \text{for } k = 2 \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots && \qquad (2) \\ p_n &= -\frac{1}{n}(S_n + p_1 S_{n-1} + p_2 S_{n-2} + \ldots + p_{n-1} S_1) && \text{for } k = n. \end{aligned}$$

Consequently, the coefficients of the characteristic polynomial $p_1, p_2, \ldots, p_n$ can be easily found when the sums $S_1, S_2, \ldots, S_n$ are known.

Thus Leverrier's scheme for seeking a characteristic determinant is the following:

(1) we calculate the powers $A^k = A^{k-1} \cdot A$ $(k = 1, 2, \ldots, n)$,

(2) we find the sums S_k which are the sums of the principal elements of the matrices A^k.

(3) and then we use formulas (2) to find the coefficients p_i $(i = 1, 2, \ldots, n)$.

The modified Leverrier method, suggested by Faddeev, consists in calculating the sequence of matrices A_1,

$A_2, \ldots, A_n$ according to the following scheme:

$$\begin{aligned}
&A_1 = A, \quad \operatorname{Tr} A_1 = q_1, \qquad B_1 = A_1 - q_1 I,\\
&A_2 = AB_1, \quad \frac{\operatorname{Tr} A_2}{2} = q_2, \qquad B_2 = A_2 - q_2 I, \qquad (3)\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&A_{n-1} = AB_{n-2}, \quad \frac{\operatorname{Tr} A_{n-1}}{n-1} = q_{n-1}, \quad B_{n-1} = A_{n-1} - q_{n-1} I,\\
&A_n = AB_{n-1}, \quad \frac{\operatorname{Tr} A_n}{n} = q_n, \quad B_n = A_n - q_n I \; (B_n \text{ is a zero matix}),\\
&q_1 = -p_1, \; q_2 = -p_2, \; \ldots, \; q_{n-1} = -p_{n-1}, \; q_n = -p_n.
\end{aligned}$$

Example 1. Use the Leverrier-Faddeev method to expand the characteristic determinant of the matrix

$$A = \begin{bmatrix} 1 & -1 & -1 & 2 \\ 2 & 3 & 0 & -4 \\ 1 & 1 & -2 & -2 \\ 1 & 1 & 0 & -1 \end{bmatrix}.$$

△ We successively obtain

$$(1)\quad A_1 = A = \begin{bmatrix} 1 & -1 & -1 & 2 \\ 2 & 3 & 0 & -4 \\ 1 & 1 & -2 & -2 \\ 1 & 1 & 0 & -1 \end{bmatrix}, \quad q_1 = \operatorname{Tr} A_1 = 1 + 3 - 2 - 1$$

$$B_1 = A_1 - q_1 I = \begin{bmatrix} 0 & -1 & -1 & 2 \\ 2 & 2 & 0 & -4 \\ 1 & 1 & -3 & -2 \\ 1 & 1 & 0 & -2 \end{bmatrix},$$

$$(2)\quad A_2 = AB_1 = \begin{bmatrix} 1 & -1 & 1 & 2 \\ 2 & 3 & 0 & -4 \\ 1 & 1 & -2 & -2 \\ 1 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & -1 & -1 & 2 \\ 2 & 2 & 0 & -4 \\ 1 & 1 & 3 & -2 \\ 1 & 1 & 0 & -2 \end{bmatrix}$$

$$= \begin{bmatrix} -1 & -2 & 2 & 4 \\ 2 & 0 & -2 & 0 \\ -2 & -3 & 5 & 6 \\ 1 & 0 & -1 & 0 \end{bmatrix}, \quad q_2 = \frac{\operatorname{Tr} A_2}{2} = \frac{-1+0+5+0}{2} = 2.$$

$$B_2 = A_2 - q_2 I = \begin{bmatrix} -3 & -2 & 2 & 4 \\ 2 & -2 & -2 & 0 \\ -2 & -3 & 3 & 6 \\ 1 & 0 & -1 & -2 \end{bmatrix}.$$

$$(3)\ A_3 = AB_2 = \begin{bmatrix} 1 & -1 & -1 & 2 \\ 2 & 3 & 0 & -4 \\ 1 & 1 & -2 & -2 \\ 1 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} -3 & -2 & 2 & 4 \\ 2 & -2 & -2 & 0 \\ -2 & -3 & 3 & 6 \\ 1 & 0 & -1 & -2 \end{bmatrix}$$

$$= \begin{bmatrix} -1 & 3 & -1 & -6 \\ -4 & -10 & 2 & 16 \\ 1 & 2 & -4 & -4 \\ -2 & -4 & 1 & 6 \end{bmatrix}, \quad q_3 = \frac{\operatorname{Tr} A_3}{3}$$

$$= \frac{-1-10-4+6}{3} =$$

$$B_3 = A_3 - q_3 I = \begin{bmatrix} 2 & 3 & -1 & -6 \\ -4 & -7 & 2 & 16 \\ 1 & 2 & -1 & -4 \\ -2 & -4 & 1 & 9 \end{bmatrix},$$

$$(4)\ A_4 = AB_3 = \begin{bmatrix} 1 & -1 & -1 & 2 \\ 2 & 3 & 0 & -4 \\ 1 & 1 & -2 & -2 \\ 1 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 3 & -1 & -6 \\ -4 & -7 & 2 & 16 \\ 1 & 2 & -1 & -4 \\ -2 & -4 & 1 & 9 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad q_4 = \frac{\operatorname{Tr} A_4}{4} = \frac{1+1+1+1}{4} = 1,$$

$$B_4 = A_4 - q_4 I = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = 0_4.$$

(5) Thus $p_1 = -q_1 = -1$, $p_2 = -q_2 = -2$, $p_3 = -q_3 = 3$, $p_4 = -q_4 = -1$ and $D(\lambda) = \lambda^4 - \lambda^3 - 2\lambda^2 + 3\lambda - 1$. ▲

The modification of Leverrier's method, suggested by Faddeev, makes it possible to find the inverse matrix A^{-1}. From formulas (3) we have $A_n = AB_{n-1}$, $B_n = A_n - q_nI = 0_n$, whence it follows that $A_n = q_nI$, or

$$AB_{n-1} = q_nI. \tag{4}$$

Premultiplying relation (4) by A^{-1}, we obtain $A^{-1}AB_{n-1} = A^{-1}q_nI$, whence we find that

$$A^{-1} = \frac{B_{n-1}}{q_n}, \text{ or } A^{-1} = \frac{B_{n-1}}{-p_n}. \tag{5}$$

Example 2. Calculate the inverse of the matrix A given in Example 1.

△ Using formula (5), we obtain

$$A^{-1}=\frac{B_3}{-p_4}=1\cdot\begin{bmatrix} 2 & 3 & -1 & -6\\ -4 & -7 & 2 & 16\\ 1 & 2 & -1 & -4\\ -2 & -4 & 1 & 9\end{bmatrix}.$$

Verification:

$$AA^{-1}=\begin{bmatrix} 1 & -1 & -1 & 2\\ 2 & 3 & 0 & -4\\ 1 & 1 & -2 & -2\\ 1 & 1 & 0 & 1\end{bmatrix}\begin{bmatrix} 2 & 3 & -1 & -6\\ -4 & -7 & 2 & 16\\ 1 & 2 & -1 & -4\\ -2 & -4 & 1 & 9\end{bmatrix}=\begin{bmatrix} 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix}. \blacktriangle$$

6.6. Using the Leverrier-Faddeev Method for Calculation of Eigenvectors

If the matrices $B_1, B_2, \ldots, B_{n-1}$ obtained by the Leverrier-Faddeev method and the roots $\lambda_1, \lambda_2, \ldots, \lambda_n$ of the characteristic polynomial $D(\lambda)$ are known, then the eigenvectors $\mathbf{x}^{(i)}$ can be found from the formula

$$\mathbf{x}^{(i)}=\lambda_i^{(n-1)}\mathbf{e}+\lambda_i^{(n-2)}\mathbf{b}_1+\lambda_i^{(n-3)}\mathbf{b}_2+\ldots+\mathbf{b}_{n-1},$$

where $\mathbf{e}$ is a unit vector and $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{n-1}$ are column vectors of the matrices $B_1, B_2, \ldots, B_{n-1}$ of the same order as $\mathbf{e}$.

Example. Calculate the eigenvectors of the matrix

$$A=\begin{bmatrix} 1 & -1 & -1 & 2\\ 2 & 3 & 0 & -4\\ 1 & 1 & -2 & -2\\ 1 & 1 & 0 & -1\end{bmatrix},$$

if the matrices B_1, B_2, B_3 (see Example 1 in 6.5) and the eigenvalues $\lambda_1=\lambda_2=1$, $\lambda_3=0.618$, $\lambda_4=-1.168$ of the characteristic polynomial $D(\lambda)=\lambda^4-\lambda^3-2\lambda^0+3\lambda-1$ are known.

△ We take $\mathbf{e}=\begin{bmatrix}0\\0\\0\\1\end{bmatrix}$ and then $\mathbf{b}_1=\begin{bmatrix}2\\-4\\-2\\-2\end{bmatrix}$, $\mathbf{b}_2=\begin{bmatrix}4\\0\\6\\-2\end{bmatrix}$ $b_3=\begin{bmatrix}-6\\16\\-4\\9\end{bmatrix}$ (the fourth columns of the matrices B_1, B_2, B_3). From the formula $\mathbf{x}^{(i)}=\lambda_i^3\mathbf{e}+\lambda_i^2\mathbf{b}_1+\lambda_i\mathbf{b}_2+\mathbf{b}_3$ we find that

$$\mathbf{x}^{(1)}=\mathbf{x}^{(2)}=1\cdot\begin{bmatrix}0\\0\\0\\1\end{bmatrix}+1\cdot\begin{bmatrix}2\\-4\\-2\\-2\end{bmatrix}+1\cdot\begin{bmatrix}4\\0\\6\\-2\end{bmatrix}+\begin{bmatrix}-6\\16\\-4\\9\end{bmatrix}=\begin{bmatrix}0\\12\\0\\6\end{bmatrix}$$

$$\mathbf{x}^{(3)}=0.618^3\begin{bmatrix}0\\0\\0\\1\end{bmatrix}+0.618^2\begin{bmatrix}2\\-4\\-2\\-2\end{bmatrix}+0.618\begin{bmatrix}4\\0\\6\\-2\end{bmatrix}$$

$$+\begin{bmatrix}-6\\16\\-4\\9\end{bmatrix}=\begin{bmatrix}-2.764\\14.472\\-1.056\\7.236\end{bmatrix},$$

$$\mathbf{x}^{(4)}=(-1.618)^3\begin{bmatrix}0\\0\\0\\1\end{bmatrix}+(-1.618)^2\begin{bmatrix}2\\-4\\-2\\-2\end{bmatrix}+(-1.618)\begin{bmatrix}4\\0\\6\\-2\end{bmatrix}$$

$$+\begin{bmatrix}-6\\16\\-4\\9\end{bmatrix}=\begin{bmatrix}-7.236\\5.528\\-18.944\\2.764\end{bmatrix}.$$

We tabulate the results of the calculations as follows:

Table 6.5

λ_i	I	II	III	IV	V	VI
$\lambda_1=\lambda_2=1$	0	2	4	−6	0	0
	0	−4	0	16	12	1
	0	−2	6	−4	0	0
	1	−2	−2	6	6	0.5
$\lambda_3=0.618$	0	0.764	2.472	−6	−2.764	0.19
	0	−1.528	0	16	14.472	1
	0	−0.764	3.708	−4	−1.056	−0.07
	0.236	−0.764	−1.236	9	7.236	0.50
$\lambda_4=-1.618$	0	5.236	−6.472	−6	−7.236	−0.38
	0	−10.472	0	16	5.528	0.29
	0	−5.236	−9.708	−4	−18.944	1
	−4.236	−5.236	3.236	9	2.764	0.15

Columns II, III and IV are the coordinates of the fourth column of the matrices B_i multiplied by the corresponding powers of λ_i and column I includes the coordinates of the vector $\lambda_i^3 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$. Column V contains the coordinates of the vectors $\mathbf{x}^{(i)}$ and column VI contains their coordinates after normalization. ▲

6.7. Danilevsky's Method

Two matrices A and B are said to be *similar* if one of them can be obtained from the other by means of a transformation with the use of a nonsingular matrix, i.e. if the equality

$$B = S^{-1} A S$$

is satisfied. If a matrix B is similar to a matrix A, then we write $B \sim A$.

In Danilevsky's method the construction of the scheme for computation is based on the principal property of similar matrices: *similar matrices have similar characteristic polynomials.*

If we use similitude transformations to reduce the matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{bmatrix} \tag{1}$$

to the so-called *Frobenius form*

$$F = \begin{bmatrix} f_{11} & f_{12} & f_{13} & \dots & f_{1,n-1} & f_{1n} \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & 1 & 0 \end{bmatrix} \tag{2}$$

and then expand the determinant

$$\det(F - \lambda I) = \begin{vmatrix} f_{11}-\lambda & f_{12} & f_{13} & \dots & f_{1,n-1} & f_{1n} \\ 1 & -\lambda & 0 & \dots & 0 & 0 \\ 0 & 1 & -\lambda & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & 1 & -\lambda \end{vmatrix} \tag{3}$$

according to the elements of the first row, we obtain

$$D(\lambda) = \det(F - \lambda I) = (f_{11} - \lambda)(-\lambda)^{n-1} - f_{12}(-\lambda)^{n-2} + f_{13}(-\lambda)^{n-3} - \dots + (-1)^{n-1} f_{1n},$$

or

$$D(\lambda) = \det(F - \lambda I) = (-1)^n (\lambda^n - p_1\lambda^{n-1} - p_2\lambda^{n-2} - p_3\lambda^{n-3} - \ldots p_n). \qquad (4)$$

Here $p_1 = f_{11}$, $p_2 = f_{12}$, $p_3 = f_{13}$, $\ldots$, $p_n = f_{1n}$ are the coefficients of the characteristic polynomial of the matrix F, which, by virtue of the similarity of the matrices F and A, are also the coefficients of the characteristic polynomial of the matrix A.

In accordance with Danilevsky's method, the matrix A can be turned into the Frobenius matrix F similar to it with the help of $n - 1$ similitude transformations which successively transform the rows of the matrix A, beginning with the last row, into the corresponding rows of the matrix F.

The scheme for transformation of the matrix A into a Frobenius matrix F similar to it. 1°. Suppose we have to transform the row $a_{n1}\ a_{n2}\ \ldots,\ a_{n,n-1}a_{nn}$ into a row $0\ 0\ \ldots\ 1\ 0$. Assuming that $a_{n,n-1} \neq 0$, we divide all the elements of the $(n - 1)$th column of the matrix A by $a_{n,\,n-1}$. Then its nth row assumes the form

$$a_{n1}a_{n2} \ldots \frac{a_{n,\,n-1}}{a_{n,\,n-1}} a_{nn}, \text{ or } a_{n1}\ a_{n2}\ \ldots\ 1\ a_{nn}.$$

2°. We subtract the $(n - 1)$th column of the transformed matrix, multiplied by the numbers a_{n1}, a_{n2}, $\ldots$, a_{nn} respectively, from all the other columns. For the nth row we obtain

$$a_{n1} - a_{n1}a_{n2} - a_{n2} \ldots 1 a_{nn} - a_{nn}, \text{ or } 0\ 0\ \ldots\ 1\ 0.$$

3°. We take the matrix M_{n-1}, obtained from the identity matrix as a result of the same transformations, as a nonsingular matrix:

$$M_{n-1} = \begin{bmatrix} 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ m_{n-1,\,1} & m_{n-1,\,2} & \cdot\ \cdot & m_{n-1,\,n-1} & m_{n-1,\,n} \\ 0 & 0 & .\,. & 0 & 1 \end{bmatrix},$$

where

$$m_{n-1,\,i} = -\frac{a_{ni}}{a_{n,\,n-1}}, \quad m_{n-1,\,n-1} = \frac{1}{a_{n,\,n-1}}. \qquad (5)$$

The operations performed are equivalent to the postmultiplication of the matrix M_{n-1} by the matrix A:

$$B = AM_{n-1} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1,\,n-1} & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2,\,n-1} & a_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n-1,\,1} & a_{n-1,\,2} & \ldots & a_{n-1,\,n-1} & a_{n-1,\,n} \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix}$$

$$\times\begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ m_{n-1,1} & m_{n-1,2} & \cdots & m_{n-1,n-1} & m_{n-1,n} \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1,n-1} & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2,n-1} & b_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ b_{n-1,1} & b_{n-1,2} & \cdots & b_{n-1,n-1} & b_{n-1,n} \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}.$$

The elements of the matrix B can be found from the formulas

$$b_{ij} = a_{ij} + a_{i,n-1} m_{n-1,j}, \quad b_{i,n-1} = a_{i,n-1} m_{n-1,n-1}. \tag{6}$$

However, the matrix $B = AM_{n-1}$ we have constructed is not similar to the matrix A.

4°. To obtain a transformation of similitude, we must premultiply the inverse matrix M_{n-1}^{-1} by the matrix B:

$$M_{n-1}^{-1} A M_{n-1} = M_{n-1}^{-1} B.$$

The inverse matrix M_{n-1}^{-1} has the form

$$M_{n-1}^{-1} = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{n,n-1} & a_{nn} \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}.$$

We set $M_{n-1}^{-1} A M_{n-1} = C$, and, consequently, $C = M_{n-1}^{-1} B$. The premultiplication of the matrix M_{n-1}^{-1} by the matrix B does not alter the transformed row of the latter and the matrix C has the form

$$C = M_{n-1}^{-1} B = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{n,n-1} & a_{nn} \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}$$

$$\times\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1,n-1} & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2,n-1} & b_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ b_{n-1,1} & b_{n-1,2} & \cdots & b_{n-1,n-1} & b_{n-1,n} \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1,n-1} & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2,n-1} & c_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ c_{n-1,1} & c_{n-1,2} & \cdots & c_{n-1,n-1} & c_{n-1,n} \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}.$$

Indeed, multiplying the matrix M_{n-1}^{-1} by B, we only change the $(n-1)$th row of the matrix B since $c_{ij} = b_{ij}$ for all the other rows. The elements of this row can be found from the formulas

$$c_{n-1,\,j} = \sum_{k=1}^{n} a_{nk} b_{kj} \quad (j = 1, 2, \ldots, n). \tag{7}$$

The matrix C obtained is similar to the matrix A and has one reduced row.

5°. Furthermore, if $c_{n-1,\,n-2} \neq 0$, we perform similar operations for the matrix C, taking the $(n-2)$th row as the principal row. Then, using the intermediate matrix $D = CM_{n-2}$, we get a matrix $I = M_{n-2}^{-1}D = M_{n-2}^{-1}CM_{n-2}$ with two reduced rows. We perform the same operations for the matrix I and so on until we obtain a Frobenius matrix.

All these transformations are written as a computational scheme. The following example shows how it is formed.

Example 1. Use Danilevsky's method to expand the characteristic determinant of the matrix

$$A = \begin{bmatrix} -4 & -3 & 1 & 1 \\ 2 & 0 & 4 & -1 \\ 1 & 1 & 2 & -2 \\ 1 & 1 & -1 & -1 \end{bmatrix}.$$

△ *1st stage.* We reduce the matrix A to the Frobenius form and compile a table to make calculations (see Table 6.6).

(1) We put the elements a_{ij} $(i, j = 1, 2, 3, 4)$ of the matrix A into rows 1-4 of the table and the control sums $a_{i5} = \sum_{j=1}^{4} a_{ij} =$ $(i = 1, 2, 3, 4)$ into the column Σ. Then we mark the element $a_{43} = -1$ which is in the third column (marked column).

(2) We write the elements of the third row of the matrix $M_{n-1} = M_3$, found from formulas (5), in row 1:

$$m_{31} = -\frac{a_{41}}{a_{43}} = -\frac{1}{-1} = 1, \quad m_{32} = -\frac{a_{42}}{a_{43}} = -\frac{1}{-1} = 1,$$

$$m_{33} = \frac{1}{a_{43}} = \frac{1}{-1} = -1, \quad m_{34} = -\frac{a_{44}}{a_{43}} = -\frac{-1}{-1} = -1,$$

$$m_{35} = -\frac{a_{45}}{a_{43}} = -\frac{0}{-1} = 0.$$

The number 0 must coincide with the sum of the elements of row I after substituting -1 for the value of the element m_{33} obtained, but in this example $m_{33} = -1$. (For the sake of convenience the number -1 is usually written beside the element m_{33} and is separated from it by a line.)

(3) We write the third row of the matrix M_3^{-1}, which must coincide with the fourth row of the original matrix A, in rows 5-8 and the column for M^{-1}.

Table 6.6

Rows	M^{-1}	Columns				Σ	Σ′
		1	2	3	4		
1		−4	−3	1	1	−5	
2		2	0	4	−1	5	
3		1	1	2	−2	2	
4		1	1	$\boxed{-1}$	−1	0	
I	M_3 M_3^{-1}	1	1	−1	−1	0	
5	1	−3	−2	−1	0	−6	−5
6	1	6	4	−4	−5	1	5
7	−1	3	3	−2	−4	0	2
8	−1	0	0	1	0	1	0
7′		0	$\boxed{-1}$	−4	−1	−6	
II	M_2 M_2^{-1}	0	−1	−4	−1	−6	
9	0	−3	2	7	2	8	6
10	−1	6	−4	−20	−9	−27	−23
11	−4	0	1	0	0	1	0
12	−1	0	0	1	0	1	0
10′		$\boxed{-6}$	0	19	9	22	
III	M_1 M_1^{-1}	0.167−1	0	3.167	1.500	3.667	
13	−6	0.500	2.000	−2.500	−2.500	−2.500	−3
14	0	1	0	0	0	1	0
15	19	0	1	0	0	1	0
16	9	0	0	1	0	1	0
13′		−3	7	24	15	43	

(4) Then we write the elements of the matrix $B = A\cdot M_3$, found from formulas (6) for the unmarked columns, in rows 5-8 and the appropriate columns.

The first column is

$$\begin{aligned} b_{11} &= a_{11} + a_{13}m_{31} = -4 + 1\cdot 1 = -3,\\ b_{21} &= a_{21} + a_{23}m_{31} = 2 + 4\cdot 1 = 6,\\ b_{31} &= a_{31} + a_{33}m_{31} = 1 + 2\cdot 1 = 3,\\ b_{41} &= a_{41} + a_{43}m_{31} = 1 + (-1)\cdot 1 = 0. \end{aligned}$$

The second column is

$$\begin{aligned} b_{12} &= a_{12} + a_{13}m_{32} = -3 + 1\cdot 1 = -2,\\ b_{22} &= a_{22} + a_{23}m_{32} = 0 + 4\cdot 1 = 4,\\ b_{32} &= a_{32} + a_{33}m_{32} = 1 + 2\cdot 1 = 3,\\ b_{42} &= a_{42} + a_{43}m_{32} = 1 + (-1)\cdot 1 = 0. \end{aligned}$$

The fourth column is

$$\begin{aligned} b_{14} &= a_{14} + a_{13}m_{34} = 1 + 1\cdot(-1) = 0,\\ b_{24} &= a_{24} + a_{23}m_{34} = -1 + 4\cdot(-1) = -5,\\ b_{34} &= a_{34} + a_{33}m_{34} = -2 + 2\cdot(-1) = -4,\\ b_{44} &= a_{44} + a_{43}m_{34} = -1 + (-1)\cdot(-1) = 0. \end{aligned}$$

The transformed elements of the third (marked) column are obtained by multiplying the initial elements by $m_{33} = -1$.

The third column is

$$\begin{aligned} b_{13} &= a_{13}m_{33} = 1\cdot(-1) = -1,\\ b_{23} &= a_{23}m_{33} = 4\cdot(-1) = -4,\\ b_{33} &= a_{33}m_{33} = 2\cdot(-1) = -2,\\ b_{43} &= a_{43}m_{33} = (-1)(-1) = 1. \end{aligned}$$

The last row of the matrix B must have the form 0 0 1 0.

To verify the calculations, we add to the matrix B the corresponding elements of the column for $\sum'$ transformed with the use of the analogous binomial formulas for $m_{35} = 0$;

$$\begin{aligned} b_{16} &= a_{15} + a_{13}m_{35} = -5 + 1\cdot 0 = -5,\\ b_{26} &= a_{25} + a_{23}m_{35} = 5 + 4\cdot 0 = 5,\\ b_{36} &= a_{35} + a_{33}m_{35} = 2 + 2\cdot 0 = 2,\\ b_{46} &= a_{45} + a_{43}m_{35} = 0 + (-1)\cdot 0 = 0. \end{aligned}$$

We write the results obtained in the column for $\sum'$ in the appropriate rows. Adding to the elements of the column for $\sum'$ the corresponding elements of the third (marked) column, we get the control sums for rows 5-8 ($i = 1, 2, 3, 4$).

The column $\sum$ is

$$b_{15} = b_{16} + a_{13} = -5 - 1 = -6,\quad b_{25} = b_{26} + a_{23} = 5-4=1,$$
$$b_{35} = b_{36} + a_{33} = 2 - 2 = 0,\quad b_{45} = b_{46} + a_{43} = 0 + 1 = 1.$$

In addition, for the sake of verification, the elements of the column $\sum$ are calculated with the use of the formula $b_{i5} = \sum_{j=1}^{4} b_{ij}$ ($i = 1, 2, 3, 4$):

$$
\begin{aligned}
b_{15} &= b_{11} + b_{12} + b_{13} + b_{14} = -3 - 2 - 1 + 0 = -6,\\
b_{25} &= b_{21} + b_{22} + b_{23} + b_{24} = 6 + 4 - 4 - 5 = 1,\\
b_{35} &= b_{31} + b_{32} + b_{33} + b_{34} = 3 + 3 - 2 - 4 = 0,\\
b_{45} &= b_{41} + b_{42} + b_{43} + b_{44} = 0 + 0 + 1 + 0 = 1.
\end{aligned}
$$

The matrix B of the form

$$
B = \begin{bmatrix} -3 & -2 & -1 & 0 \\ 6 & 4 & -4 & -5 \\ 3 & 3 & 2 & -4 \\ 0 & 0 & 1 & 0 \end{bmatrix}.
$$

(5) The transformation M_3^{-1} which is performed on the matrix B and which yields a matrix $C = M_3^{-1}B$, alters only the third row of the matrix B, i.e. the seventh row of the table. The elements of this transformed row 7′ constitute the sums of paired products of the elements of the column for M_3^{-1}, which are in rows 5-8, by the corresponding elements of each of the columns of the matrix B [see formula (7)]:

$$
\begin{aligned}
c_{31} &= 1\cdot(-3) + 1\cdot 6 + (-1)\cdot 3 + (-1)\cdot 0 = 0,\\
c_{32} &= 1\cdot(-2) + 1\cdot 4 + (-1)\cdot 3 + (-1)\cdot 0 = -1,\\
c_{33} &= 1\cdot(-1) + 1\cdot(-4) + (-1)\cdot(-2) + (-1)\cdot 1 = -4,\\
c_{34} &= 1\cdot 0 + 1\cdot(-5) + (-1)\cdot(-4) + (-1)\cdot 0 = -1.
\end{aligned}
$$

We transform the column for Σ in the same way:

$$
c_{35} = 1\cdot(-6) + 1\cdot 1 + (-1)\cdot 0 + (-1)\cdot 1 = -6.
$$

As a result we get a matrix C consisting of rows 5, 6, 7′, 8 with the control sums in the column for $\sum$:

$$
C = \begin{bmatrix} -3 & -2 & -1 & 0 \\ 6 & 4 & -4 & -5 \\ 0 & \boxed{-1} & -4 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}.
$$

The matrix C is similar to the matrix A and has one reduced row. This completes the construction of the first similitude transformation: $C = M_3^{-1}AM$.

2nd stage. Assuming the matrix C to be the original one, we separate the element $c_{32} = -1$ (the second column) and continue the process by analogy.

(1) We find the elements of the matrix $M_{n-2} = M_2$ from formulas (5):

$$m_{21} = -\frac{c_{31}}{c_{32}} = -\frac{0}{-1} = 0, \quad m_{22} = \frac{1}{c_{32}} = \frac{1}{-1} = -1,$$

$$m_{23} = -\frac{c_{33}}{c_{32}} = -\frac{-4}{-1} = -4, \quad m_{24} = -\frac{c_{34}}{c_{32}} = -\frac{-1}{-1} = -1,$$

$$m_{25} = -\frac{c_{35}}{c_{32}} = -\frac{-6}{-1} = -6.$$

We sum up $0 - 1 - 4 - 1 = -6$ ($m_{22} = -1$, if m_{22} were not equal to -1, then we should have replaced m_{22} by -1).

(2) Then we write the second row of the matrix M_2^{-1}, which coincides with the third row of the matrix C, in rows 9-12 and the column for M^{-1} (see Table 6.6). We find the elements of the matrix $D = CM_2$.

The first column is

$$d_{11} = -3 + (-2)\cdot 0 = -3, \quad d_{21} = 6 + 4\cdot 0 = 6,$$
$$d_{31} = 0 + 0 = 0.$$

The second (marked) column results from the multiplication of the corresponding elements of the matrix C by $m_{22} = -1$:

$$d_{12} = c_{12}m_{22} = (-2)\cdot(-1) = 2, \quad d_{22} = c_{22}m_{22} = 4\cdot(-1) = -4,$$
$$d_{32} = c_{32}m_{22} = (-1)\cdot(-1) = 1.$$

The third column is

$$d_{13} = c_{13} + c_{12}m_{23} = -1 + (-2)\cdot(-4) = 7,$$
$$d_{23} = c_{23} + c_{22}m_{23} = -4 + 4\cdot(-4) = -20,$$
$$d_{33} = c_{33} + c_{32}m_{23} = -4 + (-1)\cdot(-4) = 0.$$

The fourth column is

$$d_{14} = c_{14} + c_{12}m_{24} = 0\cdot 1 + (-2)\cdot(-1) = 2,$$
$$d_{24} = c_{24} + c_{22}m_{24} = -5 + 4\cdot(-1) = -9,$$
$$d_{34} = c_{34} + c_{32}m_{24} = -1 + (-1)\cdot(-1) = 0.$$

The column for Σ' is

$$d_{16} = c_{15} + c_{12}m_{25} = -6 + (-2)\cdot(-6) = 6,$$
$$d_{26} = c_{25} + c_{22}m_{25} = 1 + 4\cdot(-6) = -23,$$
$$d_{36} = c_{35} + c_{32}m_{25} = -6 + (-1)\cdot(-6) = 0.$$

The elements of the column for Σ result from the addition of the elements of the column for $\sum'$ to the corresponding elements of the marked column:

$$d_{15} = d_{16} + d_{12} = 6 + 2 = 8, \quad d_{25} = d_{26} + d_{22} = -23 - 4 = -27,$$
$$d_{35} = d_{36} + d_{32} = 0 + 1 = 1.$$

The matrix D has the form

$$D=\begin{bmatrix} -3 & 2 & 7 & 2 \\ 6 & -4 & -20 & -9 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

The transformation M_2^{-1}, which was performed on the matrix D and which yields $I = M_2^{-1}D$, alters only the second row of the matrix D, i.e. the tenth row of the table. The elements of this transformed row 10′ are the sums of the paired products of the elements of the column for M_2^{-1} which are in rows 9-12:

$$\begin{aligned}
e_{21} &= 0\cdot(-3) + (-1)\cdot 6 + (-4)\cdot 0 + (-1)\cdot 0 = -6,\\
e_{22} &= 0\cdot 2 + (-1)\cdot(-4) + (-4)\cdot 1 + (-1)\cdot 0 = 0,\\
e_{23} &= 0\cdot 7 + (-1)\cdot(-20) + (-4)\cdot 0 + (-1)\cdot 1 = 19,\\
e_{24} &= 0\cdot 2 + (-1)\cdot(-9) + (-4)\cdot 0 + (-1)\cdot 0 = 9,\\
e_{25} &= 0\cdot 8 + (-1)\cdot(-27) + (-4)\cdot 1 + (-1)\cdot 1 = 22,\\
\Sigma e_{2j} &= -6 + 0 + 19 + 9 = 22.
\end{aligned}$$

This completes the construction of the second transformation of similitude: $I = M_2^{-1}CM_2$. The matrix $I \sim C$ has two reduced rows:

$$I=\left|\begin{array}{c|ccc} -3 & 2 & 7 & 2 \\ \hline \boxed{-6} & 0 & 19 & 9 \\ \hline 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array}\right|.$$

3rd stage. We assume the matrix I to be the original one. We separate the element $e_{21} = -6$ (the first column) in it and transform the matrix I into a similar Frobenius matrix F. Continuing the process by analogy, we find, from formulas (5), the elements of the matrix $M_{n-3} = M_1$:

$$m_{11} = \frac{1}{e_{21}} = \frac{1}{-6} = -0.167,\quad m_{12} = -\frac{e_{22}}{e_{21}} = -\frac{0}{-6} = 0,$$

$$m_{13} = -\frac{e_{23}}{e_{21}} = -\frac{19}{-6} = 3.167,\quad m_{14} = -\frac{e_{24}}{e_{21}} = -\frac{9}{-6} = 1.500,$$

$$m_{15} = -\frac{e_{25}}{e_{21}} = -\frac{22}{-6} = 3.667.$$

To obtain the sum $\sum = 3.667$, we replace $m_{11} = -0.167$ by -1:

$$\sum = -1 + 0 + 3.167 + 1.500 = 3.667.$$

We write the Frobenius matrix F in rows 13-16. We first construct $G = IM_1$ and then $F = M_1^{-1}G$. In the column M_1^{-1} we write row 10′ of the matrix I (see Table 6.6).

The first (marked) column is

$$g_{11} = e_{11}m_{11} = (-3)\cdot(-0.167) = 0.500,$$
$$g_{21} = e_{21}m_{11} = (-6)\cdot(-0.167) = 1.000.$$

The second column is

$$g_{12} = e_{12} + e_{11}m_{12} = 2 + (-3)\cdot 0 = 2.000,$$
$$g_{22} = e_{22} + e_{21}m_{12} = 0 + (-6)\cdot 0 = 0.$$

The third column is

$$g_{13} = e_{13} + e_{11}m_{13} = 7 + (-3)\cdot 3.167 = -2.500,$$
$$g_{23} = e_{23} + e_{21}m_{13} = 19 + (-6)\cdot 3.167 = 0.$$

The fourth column is

$$g_{14} = e_{14} + e_{11}m_{14} = 2 + (-3)\cdot 1.500 = -2.500,$$
$$g_{24} = e_{24} + e_{21}m_{14} = 9 + (-6)\cdot 1.500 = 0.$$

The column for Σ' is

$$g_{16} = e_{15} + e_{11}m_{15} = 8 + (-3)\cdot 3.667 = -3.$$
$$g_{26} = e_{25} + e_{21}m_{15} = 22 + (-6)\cdot 3.667 = 0.$$

The column for $\sum$ is

$$g_{15} = g_{16} + g_{11} = -3 + 0.500 = -2.500$$
$$g_{15} = g_{11} + g_{12} + g_{13} + g_{14} = 0.500 + 2.000 - 2.500 - 2.500 = -2.500,$$
$$g_{25} = g_{26} + g_{21} = 0 + 1 = 1,$$
$$g_{25} = g_{21} + g_{22} + g_{23} + g_{24} = 1 + 0 + 0 + 0 = 1.$$

The elements of the transformed row 13′ are the sums of the paired products of the column for M_1^{-1} which are in rows 13-16:

$$f_{11} = (-6)\cdot 0.500 + 0\cdot 1 + 19\cdot 0 + 9\cdot 0 = -3,$$
$$f_{12} = (-6)\cdot 2.000 + 0\cdot 0 + 19\cdot 1 + 9\cdot 0 = 7,$$
$$f_{13} = (-6)\cdot(-2.500) + 0\cdot 0 + 19\cdot 0 + 9\cdot 1 = 24,$$
$$f_{14} = (-6)\cdot(-2.500) + 0\cdot 0 + 19\cdot 0 + 9\cdot 0 = 15,$$
$$\sum = (-6)\cdot(-2.500) + 0\cdot 1 + 19\cdot 1 + 9\cdot 1 = 43,$$
$$\sum = -3 + 7 + 24 + 15 = 43.$$

Thus the required Frobenius matrix F, similar to A, has the form

$$F = \begin{bmatrix} -3 & 7 & 24 & 15 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

We seek the characteristic determinant of the matrix F:

$$D(\lambda)=\det(A-\lambda I)=\det(F-\lambda I)=\begin{vmatrix} -3-\lambda & 7 & 24 & 15 \\ 1 & -\lambda & 0 & 0 \\ 0 & 1 & -\lambda & 0 \\ 0 & 0 & 1 & -\lambda \end{vmatrix}.$$

From this, expanding the determinant $D(\lambda)$ according to the elements of the first row, we obtain

$$D(\lambda)=\begin{vmatrix} -3-\lambda & 7 & 24 & 15 \\ 1 & -\lambda & 0 & 0 \\ 0 & 1 & -\lambda & 0 \\ 0 & 0 & 1 & \lambda \end{vmatrix}=(-3-\lambda)\begin{vmatrix} -\lambda & 0 & 0 \\ 1 & -\lambda & 0 \\ 0 & 1 & -\lambda \end{vmatrix}$$

$$-7\begin{vmatrix} 1 & 0 & 0 \\ 0 & -\lambda & 0 \\ 0 & 1 & -\lambda \end{vmatrix}+24\begin{vmatrix} 1 & -\lambda & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -\lambda \end{vmatrix}-15\begin{vmatrix} 1 & -\lambda & 0 \\ 0 & 1 & -\lambda \\ 0 & 0 & 1 \end{vmatrix}$$

$$=(-3-\lambda(-\lambda^3)-7\lambda^2+24(-\lambda)-15$$

$$=\lambda^4+3\lambda^3-7\lambda^2-24\lambda-15.\blacktriangle$$

Particular cases in Danilevsky's method. This method can be used without any complications if all the separated elements are nonzero (as in the example just considered). Now if as a result of the transformation of the matrix $A=[a_{ij}]$ $(i=1, 2, \ldots, n,$ $j=1, 2, \ldots, n)$ into the Frobenius matrix F, we get a matrix of the form

$$D=\begin{bmatrix} d_{11} & d_{12} & \ldots & d_{1l} & \ldots & d_{1,k-1} & d_{1k} & \ldots & d_{1,n-1} & d_{1n} \\ d_{21} & d_{22} & \ldots & d_{2l} & \ldots & d_{2,k-1} & d_{2k} & \ldots & d_{2,n-1} & d_{2n} \\ d_{l1} & d_{l2} & \ldots & d_{ll} & \ldots & d_{l,k-1} & d_{lk} & \ldots & d_{l,n-1} & d_{ln} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ d_{k1} & d_{k2} & \ldots & d_{kl} & \ldots & d_{k,k-1} & d_{kk} & \ldots & d_{k,n-1} & d_{kn} \\ 0 & 0 & \ldots & 0 & \ldots & 0 & 1 & \ldots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & 0 & \ldots & 0 & 0 & \ldots & 1 & 0 \end{bmatrix},$$

and find that $d_{k,k-1}=0$, then we cannot continue with the transformations using Danilevsky's method. Two cases are possible here.

1st case. Assume that an element of the matrix D, which lies to the left of the zero element $d_{k,k-1}$, is nonzero, say $d_{kl}\neq 0$, where $l<k-1$. Then we replace the zero element $d_{k,k-1}$ by it, i.e. interchange the $(k-1)$th and lth columns of the matrix D and simultaneously interchange its $(k-1)$th and lth rows. The new matrix is similar to the given one and we can continue with the calculations using Danilevsky's method.

2nd case. Assume that $d_{kl} = 0$ $(l = 1, 2, \ldots, k - 1)$, i.e. that the separated element and all the elements of the matrix which are to the left of the separated one are zero. Then the matrix D has the form

$$D = \left[\begin{array}{cccc|cccc}
d_{11} & d_{12} & \ldots & d_{1,k-1} & d_{1k} & \ldots & d_{1,n-1} & d_{1n} \\
d_{21} & d_{22} & \ldots & d_{2,k-1} & d_{2k} & \ldots & d_{2,n-1} & d_{2n} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
d_{k-1,1} & d_{k-1,2} & \ldots & d_{k-1,k-1} & d_{k-1,k} & \ldots & d_{k-1,n-1} & d_{k-1,n} \\
\hline
0 & 0 & \ldots & 0 & d_{kk} & \ldots & d_{k,n-1} & d_{kn} \\
0 & 0 & \ldots & 0 & 1 & \ldots & 0 & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
0 & 0 & \ldots & 0 & 0 & \ldots & 1 & 0
\end{array}\right] = \left[\begin{array}{c|c} D_1 & D_2 \\ \hline 0 & D_3 \end{array}\right].$$

We divide the matrix D into four cells so that one matrix is zero. Then the characteristic determinant $\det (D - \lambda I)$ breaks into two determinants:

$$\det (D - \lambda I) = \det (D_1 - \lambda I)\cdot\det (D_3 - \lambda I).$$

but the matrix D_3 has now the form of the Frobenius matrix and therefore it remains to reduce the matrix D_1 to this form.

Table 6.7

Rows	M^{-1}	Columns				Σ	Σ'
		1	2	3	4		
1		3	−2	1	−1	1	
2		3	−2	1	1	3	
3		5	−4	2	0	3	
4		−1	−1	$\boxed{1}$	1	0	
I	M_3 / M_3^{-1}	1	1	1 | −1	−1	0	
5	−1	4	−1	1	−2	2	1
6	−1	4	−1	1	0	4	3
7	1	7	−2	2	−2	5	3
8	1	0	0	1	0	1	0
7′		−1	0	1	0	0	

Example 2. Use Danilevsky's method to expand the characteristic determinant of the matrix

$$A=\begin{bmatrix} 3 & -2 & 1 & -1 \\ 3 & -2 & 1 & 1 \\ 5 & -4 & 2 & 0 \\ -1 & -1 & 1 & 1 \end{bmatrix}.$$

△ We tabulate the calculations (see Table 6.7). The separated element $c_{32} = 0$; we cannot continue the calculations using Dani-

Table 6.8

Rows	M^{-1}	Columns				Σ	Σ'
		1	2	3	4		
5	-1	-1	4	1	0	4	
6	-1	-1	4	1	-2	2	
7′	1	0	$\boxed{-1}$	1	0	0	
8	1	0	0	1	0	1	
II	M_2 M_2^{-1}	0	-1	1	0	0	
9	0	-1	-4	5	0	0	4
10	-1	-1	-4	5	-2	-2	2
11	1	0	1	0	0	1	0
12	0	0	0	1	0	1	0
10′		$\boxed{1}$	5	-5	2	3	
III	M_1 M_1^{-1}	1 | -1	-5	5	-2	-3	
13	1	-1	1	0	2	2	3
14	5	1	0	0	0	1	0
15	-5	0	1	0	0	1	0
16	2	0	0	1	0	1	0
13′		4	-4	2	2	4	

levsky's method. Since $r_{31} \neq 0$, we interchange the second and the first column, the first and the second row of the matrix C and continue with the calculations (see Table 6.8).

As a result we obtain a Frobenius matrix $F \backsim A$:

$$F = \begin{bmatrix} 4 & -4 & 2 & 2 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

whence we find that

$$D(\lambda) = \det(A - \lambda I) = \det(F - \lambda I) = \lambda^4 - 4\lambda^3 + 4\lambda^2 - 2\lambda + 2. \quad \blacktriangle$$

Example 3. Use Danilevsky's method to expand the characteristic determinant of the matrix

$$A = \begin{bmatrix} 0 & 1 & 3 & 2 \\ 1 & 4 & 5 & 0 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

△ We tabulate the results of the calculations (see Table 6.9). Since the separated element is equal to zero, we cannot continue the calculations using Danilevsky's scheme.

The matrix $D \backsim C$ has the form

$$D = \left[\begin{array}{c|ccc} 1 & 0.5 & -2.5 & 2.5 \\ \hline 0 & 6 & 4 & -7 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array}\right].$$

We partition the matrix D into four cells by bordering and compute $D(\lambda)$:

$$D(\lambda) = \det(D - \lambda I) = \left|\begin{array}{c|ccc} 1-\lambda & 0.5 & -2.5 & 2.5 \\ \hline 0 & 6-\lambda & 4 & -7 \\ 0 & 1 & -\lambda & 0 \\ 0 & 0 & 1 & -\lambda \end{array}\right|$$

$$= (1-\lambda) \begin{vmatrix} 6-\lambda & 4 & -7 \\ 1 & -\lambda & 0 \\ 1 & 1 & -\lambda \end{vmatrix} = (1-\lambda)\,[(6-\lambda)\,\lambda^2 + 4\lambda - 7]$$

$$= \lambda^4 - 7\lambda^3 + 2\lambda^3 + 11\lambda - 7. \quad \blacktriangle$$

Table 6.9

Rows	M^{-1}	Columns				Σ	Σ'
		1	2	3	4		
1		0	1	3	2	6	
2		1	4	5	0	10	
3		1	1	2	1	5	
4		1	1	$\boxed{1}$	1	4	
I	M_3 M_3^{-1}	−1	−1	1 −1	−1	−4	
5	1	−3	−2	3	−1	−3	−6
6	1	−4	−1	5	−5	−5	−10
7	1	−1	−1	2	−1	−1	−3
8	1	0	0	1	0	1	0
7′		−8	$\boxed{-4}$	11	−7	−8	
II	M_2 M_2^{-1}	−2	−0.25 −1	2.75	−1.75	−2	
9	−8	1	0.5	−2.5	2.5	1.5	1
10	−4	−2	0.25	2.25	−3.25	−2.75	−3
11	11	0	1	0	0	1	0
12	−7	0	0	1	0	1	0
10′		$\boxed{0}$	6	4	−7	3	

6.8. Using Danilevsky's Method for Calculation of Eigenvectors

Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ be an eigenvector of the Frobenius matrix F corresponding to the given value of λ. Then $F\mathbf{y} = \lambda\mathbf{y}$, whence we have $(F - \lambda I)\,\mathbf{y} = \mathbf{0}$,

or

$$\begin{bmatrix} f_{11}-\lambda & f_{12} & f_{13} & \cdots & f_{1n} \\ 1 & -\lambda & 0 & \ldots & -0 \\ 0 & 1 & -\lambda & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & -\lambda \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = 0. \tag{1}$$

Carrying out multiplication, we get a system for determining the coordinates $y_1, y_2, \ldots, y_n$ of the eigenvector **y**:

$$\begin{cases} (f_{11}-\lambda)\, y_1 + f_{12}y_2 + f_{13}y_3 + \cdots + f_{1n}y_n = 0, \\ y_1 - \lambda y_2 \qquad\qquad\qquad\qquad = 0, \\ \qquad y_2 - \lambda y_3 \qquad\qquad\qquad = 0, \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ \qquad\qquad\qquad y_{n-1} - \lambda y_n = 0. \end{cases} \tag{1'}$$

This system of linear equations is homogeneous. We can find its solutions to within the proportionality factor in the following way. We set $y_n = 1$ and then successively find that

$$y_{n-1} = \lambda,\ y_{n-2} = \lambda y_{n-1} - \lambda^2,\ \ldots,\ y_1 = \lambda^{n-1}.$$

Thus the required eigenvector is

$$\mathbf{y} = \begin{bmatrix} \lambda^{n-1} \\ \lambda^{n-2} \\ \vdots \\ 1 \end{bmatrix}. \tag{2}$$

Since the matrix F is similar to A, λ is also an eigenvalue of the matrix A.

We designate the eigenvector of the matrix A, corresponding to the value λ as $x = (x_1, x_2, \ldots, x_n)$. Then we obtain

$$\mathbf{x} = M_{n-1}M_{n-2} \ldots M_2M_1\mathbf{y}, \tag{3}$$

where $M_{n-1}, M_{n-2}, \ldots, M_1$ are identity matrices transformed by Danilevsky's method.

For instance, the transformation M_1, carried out on **y**, yields

$$M_1\mathbf{y} = \begin{bmatrix} m_{11} & m_{12} & \dots & m_{1n} \\ 0 & 1 & & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & 1 \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{n} m_{1k}y_k \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{k=1}^{n} m_{1k}y_k \\ \lambda^{n-2} \\ \vdots \\ 1 \end{bmatrix}.$$

Consequently, the transformation M_1 changes only the first coordinate of the vector **y**. Similarly, the transformation M_2 changes only the second coordinate of the vector $M_1\mathbf{y}$ etc. Repeating this process $n-1$ times, we get the required eigenvector **x** of the matrix A.

Example. It was shown in Example 1 of 6.7 that Danilevsky's method could be used to reduce the matrix

$$A = \begin{bmatrix} -4 & -3 & 1 & 1 \\ 2 & 0 & 4 & -1 \\ 1 & 1 & 2 & -2 \\ 1 & 1 & -1 & -1 \end{bmatrix}$$

to the Frobenius form

$$F = \begin{bmatrix} -3 & 7 & 24 & 15 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

Calculate the eigenvector $\mathbf{x}^{(1)} = (x_1, x_2, x_3, x_4)$ if $\lambda_1 = -1$.

△ We use formula (3), i.e. $\mathbf{x} = M_3M_2M_1\,\mathbf{y}$, where

$$\begin{bmatrix} \lambda^3 \\ \lambda^2 \\ \lambda \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \end{bmatrix}$$

and take the matrices M_3, M_2, M_1 from Table 6.6. We find in succession that

$$M_1\mathbf{y}=\begin{bmatrix}-0.167 & 0 & 3.167 & 1.500\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix}-1\\ 1\\ -1\\ 1\end{bmatrix}=\begin{bmatrix}-1.500\\ 1\\ -1\\ 1\end{bmatrix},$$

$$M_2M_1\mathbf{y}=\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & -1 & -4 & -1\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix}-1.5\\ 1\\ -1\\ 1\end{bmatrix}=\begin{bmatrix}-1.5\\ 2\\ -1\\ 1\end{bmatrix},$$

$$\mathbf{x}^{(1)}=M_3M_2M_1\mathbf{y}=\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 1 & 1 & -1 & -1\\ 0 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix}-1.5\\ 2\\ -1\\ 1\end{bmatrix}=\begin{bmatrix}-1.5\\ 2\\ 0.5\\ 1\end{bmatrix}.$$

We can find by analogy the eigenvectors for the values of λ_2, λ_3, λ_4 as well. ▲

6.9. Using Iterative Methods to Find the First Eigenvalue of a Matrix

Iterative methods can be used to find the first (i.e. the greatest in absolute value) eigenvalue of the matrix A without evaluating the characteristic determinant.

Assume that

$$\det (A - \lambda I) = 0 \tag{1}$$

is a characteristic equation, $\lambda_1, \lambda_2, \ldots, \lambda_n$ are its roots which are the eigenvalues of the matrix $A = [a_{ij}]$ (where $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, n$). We assume that

$$|\lambda_1| \geqslant |\lambda_2| \geqslant |\lambda_3| \geqslant \ldots \geqslant |\lambda_n|,$$

i.e. λ_1 is an eigenvalue greatest in absolute value.

Then, to find the approximate value of the root λ_1, we use the following scheme:

(1) we arbitrarily choose the initial vector $\mathbf{y}$,

(2) set up the successive iterations:

$$\begin{aligned}\mathbf{y}^{(1)} &= A\mathbf{y},\\ \mathbf{y}^{(2)} &= A\cdot A\mathbf{y} = A^2\mathbf{y},\\ \mathbf{y}^{(3)} &= A\cdot A^2\mathbf{y} = A^3\mathbf{y},\\ &\\ \mathbf{y}^{(m)} &= A\cdot A^{m-1}\mathbf{y} = A^{(m)}\mathbf{y},\\ \mathbf{y}^{(m+1)} &= A\cdot A^{(m)}\mathbf{y} = A^{(m+1)}\mathbf{y},\end{aligned}$$

(3) choose the last two values of the sequence $\mathbf{y}^m = A^m\mathbf{y}$ and $\mathbf{y}^{(m+1)} = A^{m+1}\mathbf{y}$, and then

$$\lambda_1 = \lim_{m\to\infty} \frac{y_1^{(m+1)}}{y_1^{(m)}}, \quad \text{or} \quad \lambda_1 \cong \frac{y_1^{(m+1)}}{y_1^{(m)}}, \tag{2}$$

where $y_i^{(m+1)}$ and $y_i^{(m)}$ are the respective coordinates of the vectors $\mathbf{y}^{(m+1)}$ and $\mathbf{y}^{(m)}$ $(i = 1, 2, \ldots, n)$.

Thus, taking a sufficiently large number of iteration m, we can calculate the root λ_1, greatest in absolute value, of the characteristic equation of the matrix, with any degree of accuracy. To find this root, we can use any coordinate of the vector $\mathbf{y}^{(m)}$, in particular, the arithmetic mean of the respective ratios for different coordinates.

If the choice of the initial vector $\mathbf{y}$ is poor, formula (2) may not yield a needed root or even lose sense, i.e. the limit of the ratio $y_i^{(m+1)}/y_i^{(m)}$ may not exist. It is easy to detect the latter case from the "jumping" values of the ratio. If this occurs, the initial vector must be changed. The vector $\mathbf{y}^{(m+1)}$ can be taken as the first eigenvector.

Example. Find the first eigenvalue of the matrix

$$A = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 2 & 2 \\ 0 & 1 & 1 \end{bmatrix}$$

(with three decimal digits) and the corresponding eigenvector.

△ (1) We choose the initial vector $\mathbf{y} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

(2) We set up $m = 10$ iterations:

$$\mathbf{y}^{(1)} = A\mathbf{y},\ \mathbf{y}^{(2)} = A^2\mathbf{y}^{(1)},\ \ldots,\ \mathbf{y}^{(10)} = A^{10}\mathbf{y}.$$

We tabulate the results of the calculations (see Table 6.10).

Table 6.10

$\mathbf{y}$	$A\mathbf{y}$	$A^2\mathbf{y}$	$A^3\mathbf{y}$	$A^4\mathbf{y}$	$A^5\mathbf{y}$	$A^6\mathbf{y}$	$A^7\mathbf{y}$	$A^8\mathbf{y}$	$A^9\mathbf{y}$	$A^{10}\mathbf{y}$
1	4	17	69	274	1075	4189	16 260	62 973	243 569	941 37
1	5	18	67	253	964	3693	24 193	54 650	210 663	812 58
1	2	17	25	92	345	1309	5 002	19 195	73 845	284 50

(3) Terminating the iterations with $\mathbf{y}^{(10)} = A^{10}\mathbf{y}$, we have, for different coordinates,

$$\lambda_1^{(1)} \cong \frac{y_1^{(10)}}{y_1^{(9)}} = \frac{941\,370}{243\,569} = 3.865,$$

$$\lambda_1^{(2)} \cong \frac{y_2^{(10)}}{y_2^{(9)}} = \frac{812\,585}{210\,663} = 3.857,$$

$$\lambda_1^{(3)} \cong \frac{y_3^{(10)}}{y_3^{(9)}} = \frac{284\,508}{73\,845} = 3.853.$$

(4) We calculate λ_1, as the arithmetic mean of $\lambda_1^{(1)}$, $\lambda_1^{(2)}$ and $\lambda_1^{(3)}$:

$$\lambda_1 = \frac{\lambda_1^{(1)} + \lambda_1^{(2)} + \lambda_1^{(3)}}{3} \quad \frac{3.865 + 3.857 + 3.853}{3} = 3.858.$$

(5) We can take the vector $\mathbf{y}^{(10)} = A^{10}\mathbf{y} = \begin{bmatrix} 941\,370 \\ 812\,585 \\ 284\,508 \end{bmatrix}$ as the first eigenvector of the matrix A. Normalizing it, i.e. dividing all of its coordinates by the norm of the vector equal to

$$\| \mathbf{y}^{(10)} \|_3 = \sqrt{941\,370^2 + 812\,585^2 + 284\,508^2} = 1.28 \cdot 10^6,$$

we get the first eigenvector of the matrix A corresponding to the first eigenvalue $\lambda_1 = 3.858$:

$$\mathbf{x}^{(1)} = \begin{bmatrix} 0.74 \\ 0.64 \\ 0.22 \end{bmatrix}. \blacktriangle$$

6.10. Determining the Successive Eigenvalues and the Corresponding Eigenvectors

Let the eigenvalues of the matrix A be such that

$$|\lambda_1| > |\lambda_2| > |\lambda_3| \geqslant \ldots \geqslant |\lambda_n|. \tag{1}$$

Then, to find λ_2, we can employ the so-called λ-*differences*, using the value λ_1 we have:

$$\Delta_{\lambda_1} A^m \mathbf{y} = A^{m+1}\mathbf{y} - \lambda_1 A^m \mathbf{y}, \quad \Delta_{\lambda_1} A^{m-1}\mathbf{y} = A^m \mathbf{y} - \lambda_1 A^{m-1}\mathbf{y},$$

or

$$\Delta_{\lambda_1} \mathbf{y}^{(m)} = \mathbf{y}^{(m+1)} - \lambda_1 \mathbf{y}^{(m)}, \quad \Delta_{\lambda_1} \mathbf{y}^{(m-1)} = \mathbf{y}^{(m)} - \lambda_1 \mathbf{y}^{(m-1)}, \tag{2}$$

whence, passing to the coordinates of the vectors, we find that

$$\lambda_2 \cong \frac{\Delta_{\lambda_1} y_i^{(m)}}{\Delta_{\lambda_1} y_i^{(m-1)}} = \frac{y_i^{(m+1)} - \lambda_1 y_i^{(m)}}{y_i^{(m)} - \lambda_1 y_i^{(m-1)}} \quad (i = 1, 2, \ldots, n). \tag{3}$$

Formula (3) yields rough values of λ_2 since the estimation of λ_1 was also approximate.

If the moduli of all the eigenvalues are different, then we can use formulas, similar to (3), to compute the other eigenvalues, but the successive results will be even less accurate.

The number of the iteration m in the calculation of λ_2 should be smaller than in the calculation of λ_1 to obviate the loss of accuracy when subtracting close numbers.

Example. For the matrix

$$A=\begin{bmatrix}3 & 1 & 0\\ 1 & 2 & 2\\ 0 & 1 & 1\end{bmatrix}$$

find the second eigenvalue λ_2 and the corresponding eigenvector $\mathbf{x}^{(2)}$. Carry out the calculations for $m = 8$ iterations with three decimal digits.

△ We use the table of values of $A^m\mathbf{y}$ for $m = 7, 8, 9$ (see Table 6.10).

$A^7\mathbf{y}$	$A^8\mathbf{y}$	$A^9\mathbf{y}$
16 260	62 973	243 569
14 193	54 650	210 663
5 002	19 195	73 845

We employ formula (2) to set up λ-differences:

$$\Delta_{\lambda_1}\mathbf{y}_i^{(m)}=\mathbf{y}_i^{(m+1)}-\lambda_1\mathbf{y}_i^{(m)}\quad(i=1,\ 2,\ 3).$$

For each row we take the appropriate value of λ_1: $\lambda_1^{(1)} = 3.865$, $\lambda_1^{(2)} = 3.857$, $\lambda_1^{(3)} = 3.853$. We get the following table.

Table 6.11

$A^8\mathbf{y}$	$\lambda_1 A^7\mathbf{y}$	$\Delta_{\lambda_1}A^7\mathbf{y}$	$A^9\mathbf{y}$	$\lambda_1 A^8\mathbf{y}$	$\Delta_{\lambda_1}A^8\mathbf{y}$
62 973	62 845	128	243 569	243 390	179
54 650	54 742	−92	210 663	210 785	−122
19 195	19 272	−77	73 845	73 958	−113

(3) For each row we calculate λ_2 using formula (3):

$$\lambda_2^{(1)} \cong \frac{\Delta_{\lambda_1} y_1^{(8)}}{\Delta_{\lambda_1} y_1^{(7)}} = \frac{179}{128} = 1.400, \quad \lambda_2^{(2)} \cong \frac{\Delta_{\lambda_1} y_2^{(8)}}{\Delta_{\lambda_1} y_2^{(7)}} = \frac{-122}{-92} = 1.326,$$

$$\lambda_2^{(3)} \cong \frac{\Delta_{\lambda_1} y_3^{(8)}}{\Delta_{\lambda_1} y_3^{(7)}} = \frac{-113}{-77} = 1.468.$$

(4) We determine λ_2 as the arithmetic mean of $\lambda_2^{(1)}$, $\lambda_2^{(2)}$ and $\lambda_2^{(3)}$:

$$\lambda_2 = \frac{\lambda_2^{(1)} + \lambda_2^{(2)} + \lambda_2^{(3)}}{3} = \frac{1.400 + 1.326 + 1.468}{3} = 1.398.$$

(5) As the second eigenvector we take

$$\mathbf{x}^{(2)} = \Delta_{\lambda_1} A^8 \mathbf{y} = \begin{bmatrix} 179 \\ -122 \\ -113 \end{bmatrix}.$$

Normalizing it, we have

$$\mathbf{x}^{(2)} = \begin{bmatrix} 0.73 \\ -0.50 \\ -0.46 \end{bmatrix}.$$

We can find the third eigenvalue of the given matrix knowing the trace of the matrix; since $\lambda_1 + \lambda_2 + \lambda_3 = \operatorname{Tr} A = 3 + 2 + 1 = 6$, it follows that $\lambda_3 \simeq 6 - 3.858 - 1.398 = 0.744$. ▲

Exercises

1. Use the method of direct expansion to find the characteristic polynomials of the following matrices:

$$\text{(a)}\ A = \begin{bmatrix} 2 & 5 & 7 \\ 6 & 3 & 4 \\ 5 & -2 & -3 \end{bmatrix}, \quad \text{(b)}\ A = \begin{bmatrix} 2 & -1 & 3 & -4 \\ 3 & -2 & 4 & -3 \\ 5 & -3 & -2 & 1 \\ 3 & -3 & -1 & 2 \end{bmatrix},$$

$$\text{(c)}\ A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}.$$

2. Find the characteristic numbers and eigenvectors of the matrices

$$\text{(a)}\ A = \begin{bmatrix} 3 & 4 \\ 5 & 2 \end{bmatrix}, \quad \text{(b)}\ A = \begin{bmatrix} \text{[illegible]} & 3 & -3 \\ -1 & 0 & 1 \\ 1 & 2 & -1 \end{bmatrix}.$$

3. Use Krylov's method to expand the characteristic determinants of the following matrices:

$$\text{(a) } A=\begin{bmatrix}1 & 2 & -4\\ 3 & -1 & 2\\ 1 & 0 & -1\end{bmatrix},\quad \text{(b) } A=\begin{bmatrix}2 & -2 & 1 & 3\\ 4 & -1 & -2 & -3\\ 1 & 2 & -5 & 4\\ 1 & -1 & 2 & -4\end{bmatrix},$$

$$\text{(c) } A=\begin{bmatrix}1 & -2 & 1 & -2\\ 2 & -1 & 2 & -1\\ 1 & 1 & -2 & -2\\ -2 & -2 & 1 & 1\end{bmatrix}.$$

4. Use Leverrier-Faddeev method to expand the characteristic determinants of the following matrices:

$$\text{(a) } A=\begin{bmatrix}5 & -4 & 3 & 0\\ 2 & 1 & -2 & 1\\ 1 & 1 & 1 & -1\\ 2 & 1 & -2 & 0\end{bmatrix},\quad \text{(b) } A=\begin{bmatrix}2 & -4 & 3 & -2\\ 1 & 0 & -5 & -3\\ 1 & 1 & -2 & -2\\ 0 & 1 & -1 & -1\end{bmatrix}.$$

5. Employing the Leverrier-Faddeev method, find the inverse matrices for the matrices given in Exercise 4.

6. Expand the characteristic determinant of the matrix

$$A=\begin{bmatrix}2 & -1 & 2\\ 5 & -3 & 3\\ -1 & 0 & -2\end{bmatrix}$$

and use the Leverrier-Faddeev method to calculate the eigenvector.

7. Use the iterative method to calculate the first and the second eigenvalue of the matrix

$$A=\begin{bmatrix}4 & 1 & 0\\ 1 & 2 & 1\\ 0 & 1 & 1\end{bmatrix}.$$

Chapter 7

Interpolation and Extrapolation

7.1. The Function and the Methods of Its Representation

In practical problems it is often necessary to establish relationships between processes and phenomena and to describe them by mathematical means.

Let us consider the relationships for which a certain quantity y, characterizing the process, depends on a set of unrelated quantities $x_1, x_2, \ldots, x_n$, so that every set $(x_1, x_2, \ldots, x_n)$ is associated with a single value of the quantity y. This unique correspondence between the quantity y and the set of independent variables $x_1, x_2, \ldots, x_n$ is known as a *functional dependence* and the variable y itself is a *function* of the variables $x_1, x_2, \ldots, x_n$, the formal notation being

$$y = f(x_1, x_2, \ldots, x_n).$$

Thus the expression $y = x_1^2 + 3\sqrt{x_2} + x_1 x_3^2$ is a function of three variables.

If the quantity y is a function of one independent variable x, then their relationship can be represented as

$$y = f(x).$$

For instance, the area S of a circle is a function of an independent variable, the radius R of the circle, i.e. $S = f(R)$; the specific form of this function is $S = \pi R^2$. The volume of a figure is a function of three dimensions, i.e. $V = f(x_1, x_2, x_3)$, and the form of this function depends on the shape of the figure.

Mathematical analysis gives three methods of representing functional relations: (1) analytical, (2) graphical and (3) tabular.

The **analytical method** is most convenient for representing the functional dependence $y = f(x)$ since it directly indicates actions and the sequence in which they must be performed on the variable x to obtain the corresponding value of y.

Thus, for instance, mathematical means allow us to obtain the following analytical dependence of goods and valuables supplied on credit and the seasonal expenses in agriculture on the expenses on the livestock, i.e.

$$y = 51.0203 + 0.1059\,x,$$

where y constitutes the credit against the goods and x constitutes the expenses on livestock.

Here is another example of analytical dependence: in a uniformly accelerated motion, the path traversed and the time spent are related as

$$s = vt + 0.5\,at^2.$$

The advantage of the analytical method is the possibility of obtaining the values of y for any fixed argument x with any degree of accuracy.

The drawbacks of this method are the necessity to carry out the whole sequence of computations and the lack of visuality.

These drawbacks are obviated when we use the **graphical method** of representing the function $y = f(x)$.

The *graph* of the function $y = f(x)$ is the set of points of the xy-plane whose coordinates satisfy the equation $y = f(x)$.

The **tabular method** of representing functions is widely used in engineering, physics, economy and natural sciences (and is often needed in experiments).

Assume, for instance, that as a result of an experiment we have obtained the dependence of the ohmic resistance R of a copper rod on the temperature $t°$ in the form of a table:

R	77.80	79.75	80.80	82.35	83.90	85.10
$t°$	25.0	30.1	36.0	40.0	45.1	50.0

In this experiment the value of the ohmic resistance of the copper rod varies with temperature and is a dependent variable.

The advantage of the tabular method of representing a function is that for every tabulated value of an independent variable we can immediately find, without any measurements and calculations, the corresponding value of the function.

The drawback of this method is that we cannot define the function completely, i.e. there are always values of the independent variable which are not included in the table.

7.2. Mathematical Tables

There are functions often encountered in mathematics whose calculation, despite the seeming simplicity, is very cumbersome. These functions are usually tabulated, i.e. represented as mathematical tables.

Tables of functions of one variable are especially widely used. These are tables of inverse numbers, of squares and cubes of numbers, square and cubic roots, tables of logarithms, tables of trigonometric functions, tables of exponential function and of other elementary functions. Tables of functions of two and several variables are also formed. The table of products of two numbers is an example of a table of functions of two variables.

The table is a collection of values of a function for the sequence of values of the arguments $x_0, x_1, x_2, \ldots, x_n$. It must include a collection of values of the argument such that for any values of the argument different from $x_0, x_1, x_2, \ldots, x_n$, we can obtain the value of the function with the necessary degree of accuracy.

The principal characteristics of the tables are: (1) the names of the functions whose values they express, (2) the volume, (3) the step, (4) the number of symbols of the tabulated function, (5) the number of entries.

The *name of a function* whose numerical values are tabulated is the analytical expression of that function, say, $\sin x$, $\log x$, e^x etc.

The *volume of a table* is defined by the initial and finite values of the argument. Thus, for instance, the volume of the table $y = \sin x$ envelopes the values of the argument from 0°0′ to 90°.

For almost all tabulated functions the values of the

Table 7.1

Radians	Degrees	Radians	Degrees	Degrees	Radians
(1)	(2)	(3)	(4)	(5)	(6)
0.20	11.459	0.70	40.107	20	0.34907
21	12.032	71	40.680	21	36652
22	12.605	72	41.253	22	38397
23	13.178	73	41.826	23	40143
24	13.751	74	42.399	24	41888

Continued

Degrees	Radians	Minutes	Radians	Minutes	Radians
(7)	(8)	(9)	(10)	(11)	(12)
70	1.22175	20	0.00582	50	0.01454
71	23918	21	00621	51	01484
72	25662	22	00640	52	01513
73	27409	23	00669	53	02542
74	29151	24	00698	54	01571

argument in the table form an arithmetic progression, whose common difference h is known as the *step of the table.* Thus,

$$h = x_i - x_{i-1} = \text{const } (i = 1, 2, \ldots, n).$$

Then

$$x_i = x_0 + ih \ (i = 0, 1, \ldots, n).$$

As an illustration, let us consider a fragment of the table of conversion of radians into degrees and degrees into radians (Table 7.1).

In the first two columns of the table the radian measure is an independent variable and the degrees are its function. The same is true of the third and the fourth columns. The value $h = 0.01$ radian is taken here as the step of the table.

Beginning with the fifth column the inverse of the given function is considered, where degrees (or minutes) are taken as an independent variable and the corresponding radian measure is a function of the degrees (or minutes). The step of this part of the table is equal to one degree (in the fifth and the seventh column) and to one minute (in the ninth and the eleventh column).

There are also reference tables which have a complicated two-level step. Down the column we lay off the values of the argument with a relatively large step h^* : $x_i = x_0 + ih^*$ $(i = 0, 1, \ldots, n)$ and the corresponding values of the function $y_i = f_i = f(x_i)$. Across the table, in the first row, we put the values of the argument with a finer step h, usually equal to a tenth of the large step: $h = 0.1h^*$. The second and the subsequent rows include the values of the function for an argument equal to the sum of the values x_i which are in the row and the column whose intersection is the value of the function. Thus, Table 7.2 is a fragment of the table of cubic roots from which it is easy to determine the large step down the column, $h^* = 1$, and the fine step across, $h = 0.1$.

The step of a table is usually expressed as a digit of some decimal place (less frequently as two or five digits of a certain decimal place). For instance, in the tables of squares, cubes, of square and cubic roots, in the tables of logarithms the large step $h^* = 1$, in the tables of natural logarithms and the tables of inverse numbers the large step h^* is 0.1 (see Table 7.2).

If we consider the table of sines (Table 7.3), the large step in it is one degree and the fine step is equal to six minutes.

► The next characteristic feature of a table is the *number of symbols* of the tabulated function since the values of the function $y = f(x)$ for the tabulated values of the argument in mathematical tables and the results of measurements in engineering tables are approximate values. Only valid digits of the numerical value of a function are entered in a table. This means that the error does not exceed five units of the first dropped decimal place. The values of the function for all values of x presented in the table are determined with the same absolute error. The accuracy with which the values of the function are given in the table is the *accuracy of the table*. Sometimes the accuracy is different in different parts of the table.

Table 7.2

	0	1	2	3	4	5	6	7	8	9
60	8.43433	43901	44369	48836	45303	45769	46235	46700	47165	47659
61	8.48093	48556	49018	49481	49942	50403	50954	51324	51784	52243
62	8.52702	53160	53618	54075	54542	54988	55444	55899	56354	56808
63	8.57202	57715	58168	58620	59062	59524	59975	60425	60875	61325
64	8.61774	62222	62670	63118	63566	64012	64459	64904	55350	65795

7.3

	0′	6′	12′	18′	24′	30′	36′	42′	48′	54′			1′	2′	3′	4′	5′
65°	0.9063	9070	9078	9085	9092	9100	9107	9114	9121	9128	0.9335	24°	1	2	4	5	6
66°	0.9135	9143	9150	9157	9164	9171	9178	9184	9191	9198	0.9205	23°	1	2	3	5	6
67°	0.9205	9212	9219	9225	9232	9239	9245	9252	9259	9265	0.9272	22°	1	2	3	4	6
68°	0.9272	9278	9285	9291	9298	9304	9311	9317	9323	9330	0.9336	21°	1	2	3	4	5
69°	0.9336	9342	9348	9354	9361	9367	9383	9379	9385	9391	0.9397	20°	1	2	3	4	5

In some cases, when we work with tables we must know the differences of the neighbouring values of the function given in the table, i.e. $y_{i+1} - y_i$. These differences are known as the *first-order forward and backward differences* and are sometimes written in the column of the function between its values which take part in the formation of the corresponding finite difference. The differences are written in the units of the last decimal place without zeros in front of the significant digits, and without the decimal point. For instance, in the table

x	$\sin x$	
1.000	0.84147	54
1.001	0.84201	

the finite difference $0.00054 = 0.84201 - 0.84147$ is written between the corresponding values of the function.

The *number of entries* is the next characteristic feature of a table. It is equivalent to the number of arguments of the function. Thus the tables for functional dependences $y = f(x)$ are tables with one entry. Tables 7.1, 7.2 and 7.3 given above are of this kind.

The tabulation of a function of two variables $z = f(x, y)$ leads to a table with two entries. Multiplication tables are the most widely used tables of this kind.

1	2	3	4	5	6	7	8	9
541	1082	1623	2164	2705	3246	3787	4328	4869
542	1084	1626	2168	2710	3252	3794	4336	4878
543	1085	1629	2172	2715	3258	3801	4344	4887
544	1088	1632	2176	2720	3260	3808	4352	4896
545	1090	1635	2180	2725	3270	3815	4360	4905

In Table 7.4 the three-digit multiplicand is written in the left column and one-digit multiplier is written in the upper row of the table. The step of both entries is equal to unity. To obtain the product of a three-digit number by a one-digit one, it suffices to find the row whose first column contains the multiplicand and choose the column containing the multiplier. The required product is in the row and columns obtained.

Example 1. We have to multiply 543 by 8. In Table 7.4 we find the row with 543 in it and the column numbered 8. At their intersection we read the number 4344, and this is the required product.

To multiply multi-digit numbers, we divide the multiplicand into parts containing not more than three digits and apply the technique described above to each part.

Example 2. We have to multiply 541 544 by 37. We divide the number 541 544 into two three-digit parts 541 and 544. Then we successively multiply each part by three tens and seven unities and add the resulting partial products together:

$$\begin{array}{ll} 541\times 30 = 16\,230 & 544\times 30 = 16\,320 \\ \underline{541\times 7\ \ = 3\,787} & \underline{544\times 7\ \ = 3\,808} \\ \qquad 20\,017 & \qquad 20\,128 \end{array}.$$

We place the first partial product three decimal places to the left of the second product and sum up:

$$\begin{array}{r} +\begin{array}{r} 20\,017 \\ 20\,128 \end{array} \\ \hline 20\,037\,128 \end{array}.$$

And this is the required result.

When you work with tables, please remember that there may be peculiarities in their arrangement. Therefore, when using a reference book for the first time, get acquainted with its description.

7.3. The Approximation Theory

The theory of the *approximation* of functions and its practical applications are usually needed in problem solving.

Assume, for instance, that in the process of an experiment, at the descrete moments $x_0, x_1, \ldots, x_N$ we obtained the values $f_0, f_1, \ldots, f_N$ of a quantity $f(x)$. We have to reconstruct the function $f(x)$ for other $x \neq x_i$ $(i = 0, 1, \ldots, N)$. A similar problem may arise when we repeatedly compute one composite function f at different points using a computer. Instead of doing repeated computations, it may be expedient to calculate the function f at a small number of characteristic points x_i and calculate its values at other points employing a simpler rule with the use of the values $f_i = f(x_i)$ already known.

The problems of finding the derivative $f'(x)$ and the integral $\int_a^b f(x)\,dx$ from the given values f_i can be cited as other well-known examples of approximation of functions.

Finally, the problem of approximation of functions also arises when algorithms of standard programs for calculating elementary and special functions are set up.

The classical approach to the solution of these problems requires the use of the data on hand concerning the function f in order to consider another function φ which is close, in a certain sense, to f and which admits of the requisite action on it, thus allowing the estimation of the error of such an "analytic substitution".

In the process of numerical realization of this approach, it is necessary to consider four principal problems.

1. The problem of the available information concerning the function f, i.e. the form in which the function f is given.

2. The problem of the class of the approximating functions, i.e. of the functions φ by which the function f will be approximated.

3. The problem of the closeness of the approximated and the approximating functions, i.e. of the choice of the test of goodness of fit which the function φ must satisfy.

4. The problem of the error, i.e. of the estimation of the difference of the exact and the approximate value.

There are two principal cases in the problem of the data on the function f: either the function is defined analytically or it is tabulated. The graphical method of representing a function refers either to the first or to the second case according as the specific problem posed. In what follows we shall consider, on the interval $[a, b]$, functions $f(x)$ which are continuous together with a sufficient number of their derivatives and defined by their values $f_i = f(x_i)$ at the nodal points of the given net

$$\Lambda_N\colon \{a \leqslant x_0 < x_1 < \ldots < x_N \leqslant b\}. \tag{1}$$

As concerns the class of the approximating functions, we must proceed here from two main factors. First, the approximating function must reflect the peculiar features of the function being approximated and, second, it must be convenient enough to be manipulated, i.e. admit of various actions performed on it.

In numerical analysis wide use is made of three groups of approximating functions. The first group includes functions of the form $1, x, \ldots, x^n$, whose linear combinations generate a class of all polynomials of degree not higher than n. The second group consists of trigonometric functions $\sin a_i x$ and $\cos a_i x$ which generate the Fourier series and the Fourier integral. Finally, the third group consists of exponential functions $e^{a_i x}$ which define phenomena of the type of decomposition and cumulation often encountered in reality.

Later on we shall consider in more detail the polynomial approximation, i.e. we shall take a polynomial of a degree n as the approximating function.

In this case the approximating function is usually designated as $P_n(x)$ and has the form

$$\varphi(x) \equiv P_n(x) = \sum_{k=0}^{n} a_k x^k. \tag{2}$$

The problem concerning the test of goodness of fit consists essentially in determining somehow the "distance" between the approximated and the approximating function and choosing, from the whole class of approximating functions, the function for which this "distance" is minimum.

Chebyshev's method "of fitting" is the most frequently used test of goodness of fit. It is based on the concept of the distance as the maximum value of the deviation of the function φ from the function f at the nodal points x_i:

$$\rho_1 = \max_{0 \leqslant i \leqslant N} |f(x_i) - \varphi(x_i)|. \tag{3}$$

Of especial interest is the special case when the distance $\rho_1 = 0$ for the approximating function. This means that for the tabulated function $y = f(x)$ defined by its values $y_i = f_i = f(x_i)$

x_0	x_1	...	x_N
y_0	y_1	...	y_N

we have to construct an approximating function $\varphi(x)$ which coincides, at the nodal points x_i, with the values of the given function $y = f(x)$, i.e. such that $\varphi(x_i) = y_i$.

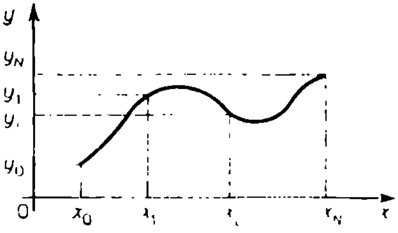


Fig. 7.1

This method of approximation, based on the test for goodness of fit of f and φ at the nodal points x_i is known as *interpolation*. If the argument x for which the approximate value of the function is sought belongs to the interval $[x_0, x_N]$, then the problem of finding the value of the function at the point x is called *interpolation in the narrow sense*. Now if the argument x is outside of the interval $[x_0, x_N]$, then the problem posed is called *extrapolation*.

In terms of geometry, the problem of interpolation for a function of one variable $y = f(x)$ means a construction on the xy-plane, of a curve which will pass through the points with coordinates (x_0, y_0), (x_1, y_1), ..., (x_N, y_N) (Fig. 7.1).

Here is one more example of the test of goodness of fit. We inroduce the concept of the distance between the functions f and φ as the sum of the squares of their deviations at the nodal points:

$$\rho = \sum^{N} [f(x_i) - \varphi(x_i)]^2. \tag{6}$$

Now we choose, as the approximating function, the function for which ρ is minimum. It is expedient to use this test when information is plentiful but is specified with a low degree of accuracy. The method of approximation based on this test is often called the **method of the least squares.** The advantages of this method are its simplicity and the order and harmony of its mathematical theory.

And now the last problem, that, concerning the accuracy of the solution obtained. In many respects it is the main problem. Indeed, in the final analysis the quality of a method depends primarily on the speed of obtaining a result with the required accuracy, or, as we say, the rate of convergence. It is therefore evident that the choice of the nodal points, of the class of approximating functions and of the test of goodness of fit must serve one purpose, the required accuracy.

At first glance the problem of the accuracy of the solution seems to be very simple: the approximate solution must differ from the exact one by not more than the specified ε. However, the question of the possibility of arbitrarily close approximation of the function f, which depends on the "parameters" enumerated above (the nodal points x_i, the class of functions φ, the test of goodness of fit of f and φ) remains open in the general case and must be investigated for every specific process of approximation.

7.4. Interpolation by Polynomials

Let us consider in more detail the problem of interpolation of the function f by algebraic polynomials.

In this case the approximating function φ is usually designated as $P_n(x)$ and has the form

$$\varphi(x) \equiv P_n(x) = \sum_{h=0}^{n} a_h x^{n-h}. \tag{1}$$

The choice of the specific value of n depends, in many respects, on the properties of the function being approximated, on the required accuracy and on the interpolation nodes. We shall see later on that the choice of the quantity n is affected essentially by the process of computation which may add an error to the result.

The condition of coincidence of f and φ at the nodal points is evidently assumed to be the test for goodness of fit.

It is natural to assume that for an unambiguous determination of $n + 1$ coefficients a_k of the polynomial P_n it is necessary to require the coincidence of f and P_n at the $(n + 1)$st nodal point:

$$f(x_i) = P_n(x_i) \quad (i = 0, 1, \ldots, n). \tag{2}$$

The polynomial $P_n(x)$, which satisfies conditions (2), is known as an *interpolating polynomial.* To emphasize the dependence of this polynomial on the function f, it is often designated as $P_n(f, x)$.

The *error of interpolation* Δ_1 in the case when it is necessary to calculate the value of the function $f(x)$ at one point x^*, is understood to be the absolute value of the difference of the exact and approximate values:

$$\Delta_1 = | f(x^*) - P_n(x^*) |. \tag{3}$$

But when the interpolation is carried out on the whole interval $[a, b]$, the maximum deviation of the polynomial P_n from the function f on the interval in question is assumed to be the error:

$$\Delta_1 = \max_{[a,b]} | f(x) - P_n(x) |.$$

Let us consider the following interpolation problem. The values $f_i = f(x_i)$ $(i = 0, 1, \ldots, n)$ of the function f are specified on the net Λ_n: $a \leqslant x_0 < x_1 < \ldots < x_n \leqslant b$ at the nodal points x_i. We have to construct an interpolating polynomial P_n which would coincide with f at the nodal points x_i and estimate the error Δ_1.

The existence and uniqueness of the interpolating polynomial follow from the theorem given below.

Theorem. *Let*: (1°) *a net* Λ_n: $a \leqslant x_0 < x_1 < \ldots < x_n \leqslant b$ *is given on the interval* $[a, b]$, *and* (2°) *arbitrary numbers* c_i $(i = 0, 1, \ldots, n)$ *are specified. Then there is a polynomial* P_n ***of degree not higher than*** n, *which assumes the assigned values* c_i *at the nodal points* x_i, *and this polynomial is unique.*

□From the conditions for determining the unknown coefficients a_k of the polynomial P_n we obtain a system of algebraic equations

$$a_0x_i^n + a_1x_i^{n-1} + \ldots + a_n = c_i \quad (i = 0, 1, \ldots, n). \qquad (4)$$

The determinant of this system

$$W = \begin{vmatrix} x_0^n & x_0^{n-1} & \ldots & x_0 & 1 \\ x_1^n & x_1^{n-1} & \ldots & x_1 & 1 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ x_n^n & x_n^{n-1} & \ldots & x_n & 1 \end{vmatrix} \qquad (5)$$

is a Vandermonde determinant which, as is known from algebra, is nonzero if the condition $x_i \neq x_j$ is satisfied for $i \neq j$. This condition is evidently fulfilled for the net Λ_n being considered. Consequently, system (4) has a unique solution (a unique collection of coefficients a_k). ■

It is evident that if we take the values f_i of the function f at the nodal points x_i as the number c_i, we shall obtain a statement on the existence and uniqueness of the interpolating polynomial $P_n(f, x)$.

The coefficients a_k of the interpolating polynomial (1) can be found by setting $c_i = f_i$ in system (4) and solving the system using, say, Cramer's rule:

$$a_k = \Delta_k / W. \qquad (6)$$

Here Δ_k is a determinant obtained from W as a result of the substitution of the column f_i of the constant terms of system (4) for the column of terms containing the $(n-k)$th power of x_i $(i = 0, 1, \ldots, n)$:

$$\Delta_k = \begin{vmatrix} x_0^n & \ldots & x_0^{n-k+1} & f_0 & x_0^{n-k-1} & \ldots & 1 \\ x_1^n & \ldots & x_1^{n-k+1} & f_1 & x_1^{n-k-1} & \ldots & 1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_n^n & \ldots & x_n^{n-k+1} & f_n & x_n^{n-k-1} & \ldots & 1 \end{vmatrix}. \qquad (7)$$

Substituting the values of a_k obtained into relation (1), we arrive at a new form of representation of the interpolating polynomial $P_n(f, x)$:

$$\begin{vmatrix} P_n & 1 & x & x^n \\ f_0 & 1 & x_0 & x_0^n \\ \ldots & & & \ldots \\ f_n & 1 & x_n & x_n^n \end{vmatrix} = 0. \qquad (8)$$

Note that in practice we usually use interpolating polynomials of the first and the second degree. And then we speak of the *linear* and *quadratic interpolation*.

Example. Using the nodal points x_0, x_1, x_2 and the corresponding values f_0, f_1, f_2 of a function, construct an interpolating polynomial representing it as a linear combination of the values f_i $(i = 0, 1, 2)$.

△ According to formula (8) we have

$$\begin{vmatrix} P_2 & 1 & x & x^2 \\ f_0 & 1 & x_0 & x_0^2 \\ f_1 & 1 & x_1 & x_1^2 \\ f_2 & 1 & x_2 & x_2^2 \end{vmatrix} = 0.$$

Expanding the determinant according to the elements of the first column, we obtain

$$P_2 \begin{vmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \end{vmatrix} - f_0 \begin{vmatrix} 1 & x & x^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \end{vmatrix} + f_1 \begin{vmatrix} 1 & x & x^2 \\ 1 & x_0 & x_0^2 \\ 1 & x_2 & x_2^2 \end{vmatrix} - f_2 \begin{vmatrix} 1 & x & x^2 \\ 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \end{vmatrix} = 0.$$

Taking into account that

$$\begin{vmatrix} 1 & x_i & x_i^2 \\ 1 & x_j & x_j^2 \\ 1 & x_k & x_k^2 \end{vmatrix} = (x_j - x_i)(x_k - x_i)(x_k - x_j),$$

we finally obtain

$$P_2(x) = f_0 \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} + f_1 \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} + f_2 \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}. \ \blacktriangle$$

7.5. The Error of Interpolation Processes

Assume that the function f is approximated by an interpolating polynomial, i.e.

$$f(x) = P_n(x) + R_n(x), \qquad (1)$$

where $R_n(x)$ is the remainder of the interpolation formula

$$f(x) \cong P_n(x). \qquad (2)$$

The remainder depends on many factors such as the properties of the function f, the interpolation parameters, the position of the point of interpolation. Therefore the study of $R_n(x)$ is a difficult problem. First of all we must estimate the error. If the point of interpolation x^* is fixed, then it is natural to take the quantity $\Delta_1 = |R_n(x^*)|$ as the estimate of the error. Now if the point x^* is unknown and the interpolation is carried out on the interval $[a, b]$, then it is expedient to take the quantity

$$\Delta_1 = \max_{[a, b]} |R_n(x)| \tag{3}$$

as estimate of the error.

Other estimates of the error can be chosen depending on a specific problem.

As a rule, the estimate of the error is sought not for an individual function but for a whole class of functions with some common properties.

We shall derive an explicit expression to estimate the error (3) of the interpolation formula (2) for the class of functions $C^{n+1}(a, b)$ which have a continuous derivative of order $n+1$ on the interval $[a, b]$.

For this purpose we prove the following theorem.

Theorem. *Assume that:* (1°) *the nodal points* x_i $(i = 0, 1, \ldots, n)$ *are distinct and belong to the interval* $[a, b]$ *together with* x^*, (2°) *the function* f *has a continuous derivative of order* $n+1$ *on* $[a, b]$. *Then there is a point* $\xi \in (a, b)$ *such that*

$$R_n(x^*) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (x^* - x_i). \tag{4}$$

□Note that if x^* coincides with one of the nodal points, then relation (4) is satisfied since its right-hand and left-hand sides are zero. Therefore we shall assume in what follows that $x^* \neq x_i$ $(i = 0, 1, \ldots, n)$. We shall consider an auxiliary function

$$\Psi(x) = f(x) - P_n(x) - k \prod_{i=0}^{n} (x - x_i), \tag{5}$$

where k is a constant chosen such that the function Ψ vanishes at $x = x^*$, i.e.

$$\Psi(x^*) = 0 = f(x^*) - P_n(x^*) - k \prod_{i=0}^{n} (x^* - x_i).$$

Hence

$$k=\frac{f(x^*)-P_n(x^*)}{\prod_{i=0}^{n}(x^*-x_i)}=\frac{R_n(x^*)}{\prod_{i=0}^{n}(x^*-x_i)}. \tag{6}$$

By virtue of such a choice of k the function Ψ vanishes on the interval $[a, b]$ at least $n+2$ times at the points x_0, $x_1, \ldots, x_n, x^*$. Then, using Rolle's theorem, we can state that in the interval (a, b) the derivative Ψ' vanishes at least $n+1$ times, the derivative Ψ'' vanishes at least n times, and so on, up to the derivative $\Psi^{(n+1)}$ which vanishes at least at one point, say, at the point $\xi\in(a, b)$.

Differentiating now the right-hand and left-hand sides of relation (5) $n+1$ times with respect to x and then setting $x=\xi$, we get zero on the left-hand side since $\Psi^{(n+1)}(\xi)=0$. The first term on the right-hand side yields the value of the derivative at the point ξ: $f^{(n+1)}(\xi)$. The second term on the right-hand side yields zero being a derivative of order $n+1$ of a polynomial of degree not higher than n. The third term is the product of the constant k by a polynomial of degree $n+1$ with the leading coefficient 1; the derivative of order $n+1$ of this polynomial is evidently equal to $(n+1)!$. Thus, summing up all we have said above, we have

$$0=f^{(n+1)}(\xi)-k(n+1)!.$$

Replacing k in this relation by its expression (6), we get the required relation (4). ■

Assume now for definiteness that

$$|f^{(n+1)}(x)|\leqslant M_{n+1},\ x\in[a, b]. \tag{7}$$

Employing this restriction and the theorem we have just proved, we arrive at the following estimate of the error for the fixed point x^*:

$$\Delta_1=|R_n(x^*)|\leqslant\frac{M_{n+1}}{(n+1)!}\prod_{i=0}^{n}|x^*-x_i|. \tag{8}$$

It is now easy to construct the estimate $|R_n|$, uniform throughout the interval $[a, b]$, for the fixed net Λ_n, namely,

$$\Delta_1=\max_{[a,b]}|R_n(x)|\leqslant\frac{M_{n+1}}{(n+1)!}\max_{[a,b]}|\omega_n(x)|, \tag{9}$$

where

$$\omega_n(x) = \prod_{i=0}^{n} (x - x_i).$$

Example 1. On the interval $[-1, 1]$ get a uniform estimate of the deviation of the function $f = 1 - \cos(\pi x/2)$ from its interpolating polynomial constructed on the nodal points $x_i = -1 + \frac{2i}{n}$ $(i = 0, 1, \ldots, n, n = 2, 3, 4)$.

△ Note first of all that for the function under consideration $M_{n+1} = (\pi/2)^{n+1}$ on the specified interval. Therefore, by virtue of estimate (9), the solution of the problem reduces to the estimation of the quantity $\max\limits_{[-1,1]} |\omega_n(x)|$. This can be done according to the ordinary rules of mathematical analysis.

1. Let us consider the case $n = 2$. Then

$$\omega_2(x) = (x+1)\,x\,(x-1), \quad \omega_2'(x) = 3x^2 - 1.$$

The roots of the polynomial $\omega_2'(x)$ are $x_{1,2} = \pm 1/\sqrt{3} \cong \pm 0.5774$. Substituting the values obtained into the expression for ω_2, we obtain

$$\max_{[-1,\,1]} |\omega_2(x)| = |\omega_2(x_1)| = 2/\sqrt{27} \cong 0.3849$$

and, consequently,

$$\Delta_1 \leqslant \left(\frac{\pi}{2}\right)^3 \cdot \frac{1}{3!} \cdot \frac{2}{\sqrt{27}} \cong 0.25.$$

2. Let now $n = 3$. In this case

$$\omega_3(x) = (x+1)\left(x+\frac{1}{3}\right)\left(x-\frac{1}{3}\right)(x-1);$$

$$\omega_3'(x) = 4x^3 - \frac{20}{9}\,x.$$

The roots of the polynomial $\omega_3'(x)$ are $x_1 = 0$ and $x_{2,3} = \pm\sqrt{5/3} \cong \pm 0.7454$. It is easy to verify that the maximum value $|\omega_3(x)|$ is attained at the points x_2 and x_3:

$$\max_{[-1,\,1]} |\omega_3(x)| = |\omega_3(x_2)| = 16/81 \cong 0.1975.$$

Therefore

$$\Delta_1 \leqslant \left(\frac{\pi}{2}\right)^4 \cdot \frac{1}{4!} \cdot \frac{16}{81} \cong 0.05.$$

3. For $n = 4$ we reduce the estimate of

$$\omega_4(x) = (x+1)\left(x+\frac{1}{2}\right) x \left(x-\frac{1}{2}\right)(x-1)$$

to the estimate obtained in item 1. Indeed, since $\omega_4(x)$ is odd, we may only find the maximum value of $|\omega_4(x)|$ on the interval $[0, 1]$. In this case

$$\max_{[0,1]} |\omega_4(x)| < 2 \cdot \frac{3}{2} \cdot \max_{[0,1]} \left| x\left(x-\frac{1}{2}\right)(x-1)\right| .$$

Using the formula $x = \frac{1}{2}(y+1)$ to change the variable on the right-hand side of the relation obtained and taking into account the results obtained in item 1, we get

$$3 \max_{[0,1]} \left| x\left(x-\frac{1}{2}\right)(x-1)\right| = \frac{3}{8} \max_{[-1,1]} |(y+1)\,y\,(y-1)|$$
$$= \frac{1}{\sqrt{48}} \cong 0.1443.$$

Thus

$$\max_{[-1,1]} |\omega_4(x)| < 1/\sqrt{48} \cong 0.1443$$

and the required estimate

$$\Lambda_1 < \left(\frac{\pi}{2}\right)^5 \cdot \frac{1}{5!} \cdot \frac{1}{\sqrt{48}} = 0.012. \ \blacktriangle$$

This example seems to substantiate the assumption that estimate (9) is practically suitable and assumes small values for the majority of functions for sufficiently large n. However, in many cases this is not so.

The matter is that only for a small class of functions (say, for entire functions) the derivatives of sufficiently high order are small, but for the majority of functions some of higher-order derivatives have a tendency of growing as $n!$. As an example let us consider the function $y = \ln x$. For this function, evidently, $y^{(n)} = (-1)^{n-1} \frac{(n-1)!}{x^n}$. Thus even in the vicinity of points where the curve $y = \log x$ seems to be smooth its derivatives of sufficiently high orders become very large and behave as $n!$.

The drawback of polynomial approximation is the absence, as a rule, of a physical meaning which usually leads to useful generalizations.

On the other hand, the simplicity and the thorough development of the theory of polynomial approximation in conjunction with the minimum of computations make this kind of approximation a convenient tool for solving various problems, all the more so as the experience of practical computations leads to good results of approximation by polynomials although either it is difficult to estimate the remainder or its estimate is too high.

We have thus considered only one aspect of the problem concerning the error, the effect of the properties of the function f on the quantity Δ_1. The problem of the dependence of the error on the arrangement of the nodal points of the net is closely connected with the properties of Chebyshev's polynomials and, therefore, we shall return to the study of this problem after we consider these polynomials. For the time being, we shall restrict our consideration to one of the possible estimates of the quantity $|\omega_n(x)|$ on the fixed net Λ_n. Let x be between x_k and x_{k+1}. We set $\max\limits_{1\leqslant i\leqslant n}(x_i-x_{i-1})=h$ and then

$$|\omega_n(x)|=\prod_{i=0}^{n}|x-x_i|<(k+1)!\,(n-k)!\,h^{n+1}\leqslant n!\,h^{n+1}. \tag{10}$$

We can therefore write inequality (8) in the form

$$\Delta_1=\max_{[a,\,b]}|R_n(x)|<\frac{M_{n+1}}{n+1}h^{n+1}. \tag{12}$$

Note that estimate (10) is rather rough and can be easily improved (do it independently as an exercise).

Example 2. With what accuracy can we calculate $\sqrt{117}$ using an interpolating polynomial for the function $y=\sqrt{x}$, taking $x_0=100$, $x_1=121$ and $x_2=144$ as the nodal points of interpolation?

$\triangle$ First of all we determine $M_3=\max\limits_{[100,\,144]}|(\sqrt{x})'''|$. To do this we find

$$y'=\frac{1}{2}x^{-1/2},\quad y''=-\frac{1}{4}x^{-3/2},\quad y'''=\frac{3}{8}x^{-5/2}.$$

Hence $M_3=\frac{3}{8}\cdot 100^{-5/2}=\frac{3}{8}\cdot 10^{-5}$. Therefore, on the basis of relation (8) we have

$$\Delta_1\leqslant\frac{3}{8}\cdot 10^{-5}\cdot\frac{1}{3!}|(117-100)(117-121)(117-144)|\cong 0.12\cdot 01^{-2}.\ \blacktriangle$$

7.6. Lagrange's Interpolating Polynomial

A direct determination of the coefficients a_h of an interpolating polynomial encounters some computational difficulties. Therefore, when we solve practical problems, we deal with special kinds of an interpolating polynomial.

In this section we consider the form of an interpolating polynomial which is known as Lagrange's form and is usually designated as $L_n(x)$. To construct L_n, we shall first consider auxiliary polynomials $l_i(x)$ of degree n which possess the following two properties:

$$l_i(x_i) = 1 \quad (i = 0, 1, \ldots, n), \tag{1}$$

$$l_i(x_k) = 0 \quad (i \neq k;\ i, k = 0, 1, \ldots, n). \tag{2}$$

These properties mean that, for instance, the polynomial $l_0(x)$ assumes a value equal to unity at the point x_0 and vanishes at the other nodal points; $l_1(x)$ assumes a value equal to unity at the point x_1 and vanishes at the other points and so on. In the general case, the polynomial $l_i(x)$ assumes a value equal to unity at the nodal point x_i and vanishes at the other points. Thus, by virtue of property (2) and the requirement that the polynomial $l_i(x)$ should be of degree n, we obtain

$$l_i(x) = c_i(x - x_0) \ldots (x - x_{i-1})(x - x_{i+1}) \ldots \\ \ldots (x - x_n). \tag{3}$$

Furthermore, using property (1), we have the following equation for determining the constant c_i:

$$l_i(x_i) = c_i(x_i - x_0) \ldots (x_i - x_{i-1}) \\ \times (x_i - x_{i+1}) \ldots (x_i - x_n) = 1.$$

From this we have

$$c_i = \frac{1}{(x_i - x_0) \ldots (x_i - x_{i-1})(x_i - x_{i+1}) \ldots (x_i - x_n)}. \tag{4}$$

We can therefore represent the explicit expression for $l_i(x)$ as follows:

$$l_i(x) = \frac{(x - x_0) \ldots (x - x_{i-1})(x - x_{i+1}) \ldots (x - x_n)}{(x_i - x_0) \ldots (x_i - x_{i-1})(x_i - x_{i+1}) \ldots (x_i - x_n)}. \tag{5}$$

Let us now set up the following linear combination of the polynomials l_i:

$$L_n(x) = \sum_{i=0}^{n} f_i l_i(x). \qquad (6)$$

Expression (6) is a polynomial of degree not higher than n. At the nodal point x_i this polynomial assumes the value f_i since the corresponding term of the sum $f_i l_i(x_i)$ is equal to f_i and the other terms of $f_j l_j(x_i)$ are zero. We have thus constructed an interpolating polynomial for the function $f(x)$. This form is known as *Lagrange's interpolating polynomial.*

Taking into account that

$$\omega_n(x) = (x - x_0)(x - x_1) \ldots (x - x_n),$$

we can consider its derivative at the point x_i:

$$\omega_n'(x_i) = (x_i - x_0) \ldots (x_i - x_{i-1})(x_i - x_{i+1}) \ldots \\ \ldots (x_i - x_n)$$

and write Lagrange's polynomial as

$$L_n(x) = \sum_{i=0}^{n} f_i \frac{\omega_n(x)}{(x - x_i)\,\omega_n'(x_i)}. \qquad (7)$$

The quantities $l_i(x)$ are the weight polynomials of the corresponding nodal points and are often called *Lagrange's multipliers.* In addition to properties (1) and (2) we shall consider one more important property of these multipliers:

$$\sum_{i=0}^{n} l_i(x) = 1. \qquad (8)$$

Indeed, assume that $f(x) \equiv 1$, and then all $f_i = 1$ $(i = 0, 1, \ldots, n)$. On the other hand, $f^{(n+1)}(x) \equiv 0$ and, by virtue of the theorem presented in 7.5, $L_n(x) = f(x) = 1$. Substituting the expression obtained into (6), we arrive at relation (8).

Example 1. Using the nodal points $x_0 = 0$, $x_1 = 1/3$ and $x_2 = 1$, construct Lagrange's interpolating polynomial for the function $f = \sin(\pi x/2)$ and get a uniform estimate of the error on the interval $[0, 1]$.

△ Note first of all that $f(x_0) = 0$, $f(x_1) = 1/2$, $f(x_2) = 1$. Then, using expression (7) for $n = 2$, we construct the required interpolating polynomial:

$$L_2(x) = \frac{1}{2}\cdot\frac{x(x-1)}{\frac{1}{3}\left(\frac{1}{3}-1\right)} + 1\cdot\frac{x\left(x-\frac{1}{3}\right)}{1\left(1-\frac{1}{3}\right)}.$$

It is easy to obtain the estimate of the error from relation (8) given in 7.5 for $n = 2$:

$$\Delta_1 \leqslant \frac{M_3}{3!}\max_{[0,\,1]}\left|x\left(x-\frac{1}{3}\right)(x-1)\right|.$$

In this case, evidently, $M_3 = (\pi/2)^3$ and $\max\limits_{[0,\,1]}\left|x\left(x-\frac{1}{3}\right)(x-1)\right| = 0.079$ and we can determine it in the same way we did it in Example 1 of 7.5. Therefore the final result is

$$\Delta_1 \leqslant \left(\frac{\pi}{2}\right)^3\cdot\frac{1}{3}\cdot 0.079 \cong 0.05. \ ▲$$

Example 2. The function $f(x)$ is tabulated as follows:

x	0	1	2	6
y	-1	-3	3	1187

Using Lagrange's interpolating polynomial, we find its value at the point $x = 4$.

△ Substituting the values of x_i and f_i for $n = 3$ and $x = 4$ into formula (7), we obtain

$$L_3(4) = -1\cdot\frac{(4-1)(4-2)(4-6)}{(-1)(-2)(-6)} - 3\cdot\frac{4(4-2)(4-6)}{1(1-2)(1-6)}$$
$$+ 3\cdot\frac{4(4-1)(4-6)}{2(2-1)(2-6)} + 1187\cdot\frac{4(4-1)(4-2)}{6(6-1)(6-2)} = 255. \ ▲$$

If we add one more point to the table in this example, we must calculate the value of the function for $x = 4$ anew. Besides this, it is seen from the example itself that the process of approximating the function with the use of Lagrange's formula is connected with long calculations. This generates a need to simplify the computations.

To make the computations easier, we compile the following table.

$x-x_0$	x_0-x_1	x_0-x_2	...	x_0-x_n	k_0
x_1-x_0	$x-x_1$	x_1-x_2	...	x_1-x_n	k_1
x_2-x_0	x_2-x_1	$x-x_2$	...	x_2-x_n	k_2
...	...	...	...	...	...
x_n-x_0	x_n-x_1	x_n-x_2	...	$x-x_n$	k_n

where $x_0, x_1, \ldots, x_n$ are the interpolation nodal points and x is the value of the argument for which we determine the approximate value using Lagrange's interpolation formula. We designate the product of the elements of the first row as k_0:

$$k_0 = (x - x_0)(x_0 - x_1) \ldots (x_0 - x_n).$$

In the general form, the product of the elements of the ith row. is

$$k_i = (x_i - x_0) \ldots (x_i - x_{i-1})(x - x_i) \times (x_i - x_{i+1}) \ldots (x_i - x_n).$$

We place the numbers $k_0, k_1, \ldots, k_n$ in the extreme right column of the table. In addition we calculate the product of the elements which lie on the principal diagonal:

$$\omega_n(x) = (x - x_0)(x - x_1) \ldots (x - x_n).$$

Then we can rewrite Lagrange's interpolating polynomial as

$$L_n(x) = \omega_n(x) \sum_{i=0}^{n} \frac{y_i}{k_i}. \tag{9}$$

Using formula (9), we solve Example 2 anew. We compile a table

4	-1	-2	-6	-48
1	$4-1$	$1-2$	$1-6$	15
2	$2-1$	$4-2$	$2-6$	-16
6	$6-1$	$6-2$	$4-6$	-240

and find that $\omega_3(4) = -48$. The approximate value of the function at the point $x = 4$, i.e. $f(4) \simeq L_3(4)$, can be found from the formula

$$L_3(x) = \omega_3(x) \sum_{i=0}^{3} \frac{y_i}{k_i},$$

or

$$L_3(4) = -48\left[\frac{-1}{-48} + \frac{-3}{15} + \frac{3}{-16} + \frac{1187}{-240}\right] = -255. \ \blacktriangle$$

Lagrange's interpolation formula is noticeably simpler when the interpolation nodal points are *equispaced*, i.e. $h = x_{i+1} - x_i = \text{const}$, where h is the step of interpolation. We introduce the designation $t = (x - x_0)/h$. From formula (5) we have

$$l_i(x) = \frac{(x-x_0)\dots(x-x_{i-1})(x-x_{i+1})\dots(x-x_n)}{(x_i-x_0)\dots(x_i-x_{i-1})(x_i-x_{i+1})\dots(x_i-x_n)}.$$

Since

$$\begin{aligned} x - x_0 &= th, \\ x - x_1 &= th - h = h(t-1), \\ &\dots\dots\dots\dots\dots \\ x - x_i &= th - ih = h(t-i), \\ &\dots\dots\dots\dots\dots \\ x - x_n &= th - nh = h(t-n), \end{aligned}$$

it follows that

$$l_i = \frac{t(t-1)\dots(t-i+1)(t-i-1)\dots(t-n)h^n}{ih(i-1)h\dots1h(-1)h\dots[-(n-i)h]}. \qquad (10)$$

Note that a part of the product in the denominator is

$$ih(i-1)h\dots h = i!\,h^i$$

and the other part is

$$(-h)\dots[-(n-i)h] = (-1)^{n-i}(n-i)!\,h^{n-i}.$$

Multiplying the numerator and denominator on the right hand side of relation (10) by $(-1)^{n-i}(t-i)$, we obtain

$$\begin{aligned} l_i &= \frac{t(t-1)\dots(t-n)}{(t-i)\,i!\,(n-i)!}(-1)^{n-i} \\ &= (-1)^{n-i}\frac{\binom{n}{i}}{t-i}\,\frac{t(t-1)\dots(t-n)}{n!}, \end{aligned}$$

where $\binom{n}{i} = \dfrac{n!}{i!\,(n-i)!}$.

Then Lagrange's interpolating polynomial for the equispaced nodal points of interpolation can be written as

$$L_n(x) = L_n(x_0 + ht) \frac{t(t-1)\dots(t-n)}{n!}$$

$$\times \sum_{i=0}^{n} (-1)^{n-i} \frac{\binom{n}{i}}{t-i} y_i. \qquad (11)$$

Example 3. The function $y = \sin x$ is tabulated as

x	0	$\pi/4$	$\pi/2$
y	0	0.707	1

Using Lagrange's interpolating polynomial, find its value at the point $x^* = \pi/6$. Estimate the error Λ_1.

△ We first find $t^* = \left(\frac{\pi}{6} - 0\right)\left(\frac{\pi}{4}\right)^{-1} = \frac{2}{3}$. Substituting the value of t^* we have obtained and the values of y_i for $n = 2$ into formula (11), we have

$$L_2\left(\frac{\pi}{6}\right) = \frac{(2/3)(2/3-1)(2/3-2)}{2!}$$

$$\times\left(\frac{2}{2/3-0}\cdot 0 - \frac{2}{2/3-1}\cdot 0.707 + \frac{1}{2/3-2}\cdot 1\right) = 0.517.$$

To estimate the error, we use formula (8) from 7.5. Here $M_3 = \max\limits_{[0,\pi/2]} |(\sin x)'''| = 1$, $x^* = \pi/6$, and therefore

$$\Lambda_1 = \frac{1}{3!}\cdot\frac{\pi}{6}\cdot\frac{\pi}{12}\cdot\frac{\pi}{3} = 0.024.$$

Note that when calculating the error we must turn the degree measure into a radian one.

Thus, rounding off the result to two decimal places, we get $\sin(\pi/6) = 0.52 \pm 0.03$. ▲

7.7. Finite Differences

In the preceding sections we considered various problems of the interpolation theory on an arbitrary net of nodal points. In practical computations the information about the function (in the form of its values) is often given on a uniform net, i.e. for equispaced nodal points. In this case, not only the forms of the interpolation polynomials become simpler, but the computation process is diminished, and this is a factor of great significance in practical computations. When constructing interpolating polynomials on a uniform net, we use quantities known as *finite differences.*

Consider a uniform net with a step h: $x_i = x_0 + ih$ $(i = 0, \pm 1, \pm 2, \dots)$, at whose nodal points the values of $f_i = f(x_i)$ of the function $f(x)$ are given.

Three types of finite differences can be encountered in mathematical literature: *forward differences* $\Delta^k f_i$, *backward differences* $\nabla^k f_i$ and *central differences* $\delta^k f_i = f^k_i$.

A *finite difference of the first order* is a difference between the values of the function at the given nodal point and at the preceding one:

$$\begin{gathered} f_1 - f_0 = \Delta f_0 = \nabla f_1 = \delta f_{1/2} = f^1_{1/2}, \\ f_2 - f_1 = \Delta f_1 = \nabla f_2 = \delta f_{3/2} = f^1_{3/2}, \\ \dots\dots\dots\dots \\ f_{i+1} - f_i = \Delta f_i = \nabla f_{i+1} = \delta f_{i+1/2} = f^1_{i+1/2}. \end{gathered} \tag{1}$$

This definition can also be written as

$$\Delta f_i = f_{i+1} - f_i, \quad \nabla f_i = f_i - f_{i-1}, \quad \delta f_i = f_{i+1/2} - f_{i-1/2}. \tag{2}$$

A *finite difference of the second order* is a difference between the values of the first finite difference at a given nodal point and the preceding one:

$$\begin{gathered} \Delta^2 f_i = \Delta f_{i+1} - \Delta f_i, \quad \nabla^2 f_i = \nabla f_i - \nabla f_{i-1}, \\ \delta^2 f_i = f^2_i = \delta f_{i+1/2} - \delta f_{i-1/2}. \end{gathered} \tag{3}$$

Finite differences of an arbitrary order k are defined in a similar way:

$$\begin{gathered} \Delta^k f_i = \Delta^{k-1} f_{i+1} - \Delta^{k-1} f_i; \\ \nabla^k f_i = \nabla^{k-1} f_i - \nabla^{k-1} f_{i-1}, \\ \delta^k f_i = f^k_i = \delta^{k-1} f_{i+1/2} - \delta^{k-1} f_{i-1/2}. \end{gathered} \tag{4}$$

In some interpolation formulas considered below, in addition to differences (4) we use the arithmetic means of the adjacent finite differences of the same order:

$$\mu f_i^k = \frac{1}{2}(f_{i+1/2}^k + f_{i-1/2}^k), \quad \mu f_{i+1/2}^k = \frac{1}{2}(f_{i+1}^k + f_i^k). \tag{5}$$

The first of these quantities is used for an odd k and the second, for an even k.

It is convenient to write the finite differences of the function f as tables. In that case the backward and forward finite differences are written as horizontal tables (Tables 7.5 and 7.6) and central finite differences as central tables (Table 7.7).

Let us consider some properties of finite differences.

Table 7.5

x	f	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$
x_0	f_0	Δf_0	$\Delta^2 f_0$	$\Delta^3 f_0$	$\Delta^4 f_0$
x_0+h	f_1	Δf_1	$\Delta^2 f_1$	$\Delta^3 f_1$	...
x_0+2h	f_2	Δf_2	$\Delta^2 f_2$	...	
x_0+3h	f_3	Δf_3	...		
x_0+4h	f_4	...			
...	...	...	...	...	...

Table 7.6

x	f	∇f	$\nabla^2 f$	$\nabla^3 f$	$\nabla^4 f$
...	...	...	...	...	...
x_0-4h	f_{-4}	...			
x_0-3h	f_{-3}	∇f_{-3}	...		
x_0-2h	f_{-2}	∇f_{-2}	$\nabla^2 f_{-2}$	...	
x_0-h	f_{-1}	∇f_{-1}	$\nabla^2 f_{-1}$	$\nabla^3 f_{-1}$	...
x_0	f_0	∇f_0	$\nabla^2 f_0$	$\nabla^3 f_0$	$\nabla^4 f_0$

Table 7.7

x	f	δf	$\delta^2 f$	$\delta^3 f$	$\delta^4 f$
...	...	...	...	...	...
x_0-2h	f_{-2}				
		$\delta f_{-3/2}$			
x_0-h	f_{-1}		$\delta^2 f_{-1}$		
		$\delta f_{-1/2}$		$\delta^3 f_{-1/2}$	
x_0	f_0		$\delta^2 f_0$		$\delta^4 f_0$
		$\delta f_{1/2}$		$\delta^3 f_{1/2}$	
x_0+h	f_1		$\delta^2 f_1$		
		$\delta f_{3/2}$			
x_0+2h	f_2				
...	...	...	...	...	...

1°. *Forward, backward and central differences are connected by the relations*

$$\Delta^k f_i = \nabla^k f_{i+k} = \delta^k f_{i+k/2}, \qquad (6)$$

which we can easily prove by induction proceeding from the definition of the finite differences (4).

□ For $k = 1$ relations (6) are obvious since, by virtue of relations (4), we have

$$\Delta f_i = f_{i+1} - f_i, \quad \nabla f_{i+1} = f_{i+1} - f_i, \quad \delta f_{i+1/2} = f_{i+1} - f_i.$$

Let now relations (6) hold true for any $k \leqslant m - 1$. We shall show that in this case they are also valid for $k = m$ and, consequently, for all k. Using relations (4) and the assumption on the validity of (6) for $k \leqslant m - 1$, we have

$$\begin{aligned} \Delta^m f_i &= \Delta^{m-1} f_{i+1} - \Delta^{m-1} f_i \\ &= \nabla^{m-1} f_{i+m} - \nabla^{m-1} f_{i+m-1} = \nabla^m f_{i+m}. \end{aligned}$$

$$\begin{aligned}\Delta^m f_i &= \Delta^{m-1} f_{i+1} - \Delta^{m-1} f_i \\ &= \delta^{m-1} f_{i+\frac{m+1}{2}} - \delta^{m-1} f_{i+\frac{m-1}{2}} = \delta^m f_{i+\frac{m}{2}}. \ \blacksquare\end{aligned}$$

2°. *The finite difference satisfies the equality*

$$\Delta\,(af + bg)_i = a\,\Delta f_i + b\,\Delta g_i, \tag{7}$$

where a and b are constants.

□Indeed,

$$\begin{aligned}\Delta\,(af + bg)_i &= af_{i+1} + bg_{i+1} - (af_i + bg_i) \\ &= a\,\Delta f_i + b\Delta g_i. \ \blacksquare\end{aligned}$$

This property means, in particular, that the finite difference of the sum or the difference of two functions is equal to the sum or the difference of the finite differences of those functions respectively and also that the finite difference of the product of a function by a constant factor is equal to the product of that factor by the finite difference of the function.

3°. *A finite difference is connected with the corresponding derivative by a relation*

$$\Delta^k f_i = h^k f^{(k)}(\xi), \quad \xi \in (x_i,\ x_i + kh). \tag{8}$$

A consequence of relation (8) is that finite differences of order n of a polynomial of degree n are constant and equal to $h^n n! a_0$ and finite differences of any higher order are equal to zero (a_0 is a coefficient of the polynomial in the highest degree of x).

4°. *A finite difference of order k can be represented as the following linear combination of the values of* f_i:

$$\Delta^k f_i = \sum_{j=0}^{k} (-1)^j \binom{k}{j} f_{i+k-j}, \tag{9}$$

where $\binom{k}{j} = \frac{k!}{j!\,(k-j)!}$ is the number of combinations of k elements taken j at a time (with $0! = 1$).

□We use induction. For $k = 1$ this relation is obvious since it is a definition of the first finite difference: $\Delta f_i = f_{i+1} - f_i$.

Let now relation (9) hold true for some $k = m$. Then

$$\Delta^{m+1} f_i = \Delta^m f_{i+1} - \Delta^m f_i$$

$$= \sum_{j=0}^{m} (-1)^j \binom{m}{j} f_{i+1+m-j} - \sum_{j=0}^{m} (-1)^j \binom{m}{j} f_{i+m-j}$$

$$= \binom{m}{0} f_{i+1+m} + \sum_{j=1}^{m} (-1)^j \binom{m}{j} + \binom{m}{j-1} f_{i+1+m-j}$$

$$+ (-1)^{m+1} \binom{m+1}{m+1} f_i = \sum_{j=0}^{m+1} (-1)^j \binom{m+1}{j} f_{i+m+1-j}.$$

We have used the properties of combinations: $\binom{m}{j-1} = \binom{m+1}{j} - \binom{m}{j}$ and $\binom{m}{0} = \binom{m+1}{0} = \binom{m}{m} = \binom{m+1}{m+1} = 1$. Thus if relation (9) holds true for $k = m$, then it also holds true for $k = m + 1$. ■

Example 1. Compile a horizontal table of the finite differences of the function $y = x^3 + 3x^2 - x - 1$ proceeding from the initial value $x = 0$ and taking $h = 1$ as a step.

△ Setting $x_0 = 0$, $x_1 = 1$, $x_2 = 2$, . . ., we find the corresponding values:

x	0	1	2	3	4	5	. . .
y	−1	2	17	50	107	194	. . .

We seek the finite differences of the first order:

$$\Delta y_0 = y_1 - y_0 = 2 - (-1) = 3,$$
$$\Delta y_1 = y_2 - y_1 = 17 - 2 = 15,$$
$$\Delta y_2 = y_3 - y_2 = 50 - 17 = 33,$$
$$\Delta y_3 = y_4 - y_3 = 107 - 50 = 57,$$
$$\Delta y_4 = y_5 - y_4 = 194 - 107 = 87, \ldots$$

Next we seek the finite differences of the second order:

$$\Delta^2 y_0 = \Delta y_1 - \Delta y_0 = 15 - 3 = 12,$$
$$\Delta^2 y_1 = \Delta y_2 - \Delta y_1 = 33 - 15 = 18,$$
$$\Delta^2 y_2 = \Delta y_3 - \Delta y_2 = 57 - 33 = 24,$$
$$\Delta^2 y_3 = \Delta y_4 - \Delta y_3 = 87 - 57 = 30, \ldots$$

And then we find the finite differences of the third order:

$$\Delta^3 y_0 = \Delta^2 y_1 - \Delta^2 y_0 = 18 - 12 = 6,$$
$$\Delta^3 y_1 = \Delta^2 y_2 - \Delta^2 y_1 = 24 - 18 = 6,$$
$$\Delta^3 y_2 = \Delta^2 y_3 - \Delta^2 y_2 = 30 - 24 = 6, \ldots$$

We see that the third finite differences $\Delta^3 y_0$, $\Delta^3 y_1$, $\Delta^3 y_2$ are constant. This can be explained by the fact that the function $f(x)$ is a third-degree polynomial. The third finite difference can also be found from the formula

$$\Delta^n P_n(x) = n!\ a_0 h^n,$$

i.e. $\Delta^3 P_3(x) = 3! \cdot 1 \cdot 1^3 = 6$, and the finite differences of the fourth order are equal to zero.

We compile a table of finite differences:

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
0	−1	3	12	6	0
1	2	15	18	6	0
2	17	33	24	6	
3	50	57	30		
4	107	87			
5	194				

In what follows, it is expedient to tabulate finite differences encountered in calculations. ▲

Since the initial values f_i of the function are given, as a rule, with a certain error ε which is the rounding errors or random errors, it is advisable to consider the effect these factors exert on the errors of higher-order finite differences.

We begin with the influence exerted by random errors and the rounding errors. Assume that we have got $f_i^k + \varepsilon$ instead of f_i^k. Then the table of finite differences assumes the form shown on p. 315 (see Table 7.8).

By virtue of relation (9) this means that the error ε in the difference of order k is extended to the difference of order $k+m$ with coefficients $(-1)^j \binom{m}{j}$.

Now if all the initial values f_i are given with the same error ε, then this error is extended to the differences of order m with a coefficient 2^m and grows rapidly with an increase in m $(\Delta (f_i^m) = 2^m \varepsilon)$.

If the derivatives of sufficiently high orders of the function f remain bounded, then it follows from formula (8)

Table 7.8

...	...	...	...
f^h_{i-2}			
	$f^{h+1}_{i-3/2}$		
f^h_{i-1}		$f^{h+2}_{i-1}+\varepsilon$	
	$f^{h+1}_{i-1/2}-\varepsilon$		$f^{h+3}_{i-1/2}-3\varepsilon$
$f^h_i+\varepsilon$		$f^{h+2}_i-2\varepsilon$	
	$f^{h+1}_{i+1/2}-\varepsilon$		$f^{h+3}_{i+1/2}+3\varepsilon$
f^h_{i+1}		$f^{h+2}_{i+1}+\varepsilon$	
	$f^{h+1}_{i+3/2}$		
f^h_{i+2}			
...			

that the corresponding finite differences f_i^m decrease with an increase in m. Therefore there comes a point where the errors of the finite differences resulting either from rounding off or from the errors in the initial data, become comparable with the values of the finite differences themselves or even exceed them. Consequently, the data in the table of these differences will be, in essense, the data on the differences of the errors and not the data on the function and it will not be expedient to use it. In that case we say that the order of the last finite differences which can still be used in calculations is the *order of correctness of the table of finite differences.*

Example 2. Consider a table of values of the function $f = \sin x$:

x	46°	47°	48°	49°	50°	51°	52°
f	0.7198	0.7314	0.7431	0.7547	0.7660	0.7771	0.7880

All the digits in the table are correct in the narrow sense. Compile a table of finite differences and find the order of correctness of the table.

△ We calculate the finite differences and compile a table.

Table 7.9

x	f	f^1	f^2	f^3	f^4	f^5	f^6
46°	0.7193						
		121					
47°	0.7314		−4				
		117		3			
48°	0.7431		−1		−5		
		116		2		8	
49°	0.7547		−3		3		−12
		113		1		−4	
50°	0.7660		−2		−1		
		111		0			
51°	0.7771		−2				
		109					
52°	0.7880						
	0.00005	1	2	4	8	16	32

Note that it is customary to write finite differences in the units of the last decimal place of the values of the function.

The last row of the table contains the corresponding absolute errors. Evidently, the absolute values of the third finite differences are comparable with their error and the absolute values of the subsequent differences are essentially smaller than their errors. Therefore the order of correctness of Table 7.9 is 2. ▲

The fact that it is inexpedient to use finite differences of an order higher than two in the example given above can also be substantiated as follows. Since the initial data on f_i are given with an error $\varepsilon = 0.5 \cdot 10^{-4}$, then it is not advisable to take into consideration the finite differences which are smaller in absolute value than the errors. On the other hand, using relation (8), we get $\Delta^k f_i \cong (\pi/180)^k$ and, therefore, the order of correctness of Table 7.9 is defined by the following inequalities:

$$(\pi/180)^k > 0.5 \cdot 10^{-4} \geqslant (\pi/180)^{k+1},$$

which are evidently satisfied for $k = 2$.

In the general case, the order of correctness of a table of finite differences in this sense is the least value of k satisfying the inequality

$$h^{k+1} M_{k+1} \leqslant \varepsilon, \text{ where } M_\nu = \max_{[x_i, x_{i+\nu}]} |f^{(\nu)}(x)|. \qquad (10)$$

We have discussed the influence exerted by the error of the initial data on the degree of the interpolating polynomial. Besides this, if the values f_i are approximate or, for some reason, the calculation of the value of the polynomial $P_n(x^*)$ cannot be absolutely accurate, then, in fact, we get only an approximate value $\overline{P}_n(x^*)$ for an exact $P_n(x^*)$. In that case the error of calculation of $\Delta_2(\overline{P}_n) = |P_n(x^*) - \overline{P}_n(x^*)|$ is evaluated according to the general rules of calculation of an error of a function.

Let us consider a Lagrange polynomial $L_n(x) = \sum_{i=0}^{n} f_i l_i(x)$, for example. Assume that we have to calculate $L_n(x^*)$ for the given values f_i and their errors ε_i. The values of the Lagrange coefficients $l_i(x)$ have been tabulated for equispaced nodal points and we may consider them to be exact numbers since they have been obtained from the exact values of the nodal points and the exact x^*. Therefore we have the following inequality for the Lagrange polynomial:

$$\Delta_2(\overline{L}_n) < \sum_{i=0}^{n} \varepsilon_i \, |l_i(x^*)|.$$

In the case when all ε_i are the same and equal to ε, we get

$$\Delta_2(\overline{L}_n) \leqslant \varepsilon \sum_{i=0}^{n} |l_i(x^*)|.$$

The calculation error for other forms of an interpolating polynomial can be found in a similar fashion.

Example 3. On the interval $[-1, 1]$ get a uniform estimate of the calculation error of the values of a Lagrange interpolating polynomial constructed for the function $f = \cos(\pi x/2)$ using the nodal points $x_0 = -1/2$, $x_1 = 0$, $x_2 = 1/2$.

△ Since $f_0 = f_2 = 1/\sqrt{2} = 0.707 \pm 0.0002$ and $f_1 = 1$ is an exact number, the required computational error has the form

$$\Delta_2(\bar{L}_2) = 0.0002\left|\frac{x^*\left(x^*-\frac{1}{2}\right)}{-\frac{1}{2}\left(-\frac{1}{2}-\frac{1}{2}\right)}\right| + 0.0002\left|\frac{\left(x^*+\frac{1}{2}\right)x^*}{\left(\frac{1}{2}+\frac{1}{2}\right)\cdot\frac{1}{2}}\right|$$

$$= 0.0004\left[\left|x^*\left(x^*-\frac{1}{2}\right)\right| + \left|\left(x^*+\frac{1}{2}\right)x^*\right|\right].$$

It is easy to show that on the interval $[-1, 1]$ the quantity $\Delta_2(\bar{L}_2)$ assumes the maximum value at the points $x^* = \pm 1$ and, therefore, the required estimate is $\Delta_2(\bar{L}_2) = 0.0008$. ▲

7.8. Stirling and Bessel Interpolating Polynomials

We shall first dwell on the problem of choosing interpolation nodes for a fixed degree of a polynomial which is significant from the point of view of an interpolation error. Let us consider an expression for the remainder of the interpolating polynomial:

$$\frac{f^{(n+1)}}{(n+1)!}\prod_{i=0}^{n}(x^*-x_i), \quad \xi\in(x_0, x_n).$$

Since an interpolation interval is usually not large, the derivative $f^{(n+1)}(x)$ has a small range of variation. Consequently, the range of variation of the value of the error is defined, in the main, by the product

$$|\omega_n(x^*)| = \prod_{i=0}^{n} |x^*-x_i|.$$

This value is minimum if the nodes closest to x^* are taken as the interpolation nodes. Thus, for an even degree $n = 2k$ of the interpolating polynomial we must take a node closest to the point x^* and k nodes on the left and k nodes on the right of it, and for an odd degree $n = 2k + 1$ we must take $k + 1$ nodes on the left and $k + 1$ nodes on the right of the point x^*.

Let us now construct an interpolating polynomial for the function f specified by its values f_i at the nodes x_i of a uniform net with a step h.

Let the point x^* lie close to a certain node which we denote by x_0. We have to construct an interpolating polynomial of an even degree. In accordance with what we have said, we must take, as interpolation nodes, a net which is symmetric with respect to the node x_0.

$$\ldots, x_{-k}, \ldots, x_{-1}, x_0, x_1, \ldots, x_k, \ldots$$

We introduce a new variable t which serves to transfer the reference point to the point x_0:

$$t = (x - x_0)/h. \tag{1}$$

In this case $t^* = (x^* - x_0)/h$.

We construct an interpolating polynomial in the following form:

$$\begin{aligned} P_{2k}(x) = P_{2k}(x_0 + th) = a_0 + \frac{a_1}{1!}t + \frac{a_2}{2!}t^2 + \\ \ldots + \frac{a_{2k-1}}{(2k-1)!}t(t^2-1^2)\ldots(t^2-(k-1)^2) \\ + \frac{a_{2k}}{(2k)!}t^2(t^2-1^2)\ldots(t^2-(k-1)^2). \end{aligned} \tag{2}$$

The unknown coefficients a_i can be found from the conditions for coincidence of the polynomial P_{2k} and the function f at the nodes x_i. We note that relation (1) puts the quantity $t = i$ into correspondence with each of the nodes x_i. For instance, $t = 0$ is associated with the node x_0 and $t = -3$ with the node x_{-3}.

Thus, to find the coefficients a_i, we get a system of linear equations

$$P_{2k}(x_0 + ih) = f_i \quad (i = 0, \pm 1, \ldots, \pm k). \tag{3}$$

The structure of this system is such that a_0 can be found directly from the first equation of system (3):

$$P_{2k}(x_0) = a_0 = f_0$$

and the search for the other coefficients reduces to the successive solution of systems of two equations in two unknowns:

$$P_{2k}(x_0 + h) = f_0 + a_1 + \frac{a_2}{2} = f_1,$$

$$P_{2k}(x_0 - h) = f_0 - a_1 + \frac{a_2}{2} = f_{-1}.$$

Hence (see the designations in 7.7) $a_1 = \mu f_0^1$, $a_2 = f_0^2$.

Continuing this process, we note that at the jth step the determinant of the system of equations relative to the coefficients a_{2j-1} and a_{2j} has the form

$$\begin{vmatrix} 1 & 1/2 \\ -1 & 1/2 \end{vmatrix} = 1 \neq 0.$$

Consequently, all the coefficients a_i of polynomial (2) are uniquely defined by the system of equations (3) and, by virtue of the theorem on uniqueness, expression (2) is an interpolating polynomial for the function f.

As a result of simple transformations we arrive at the following expressions for the coefficients:

$$a_0 = f_0,\quad a_{2j-1} = \mu f_0^{2j-1},\quad a_{2j} = f_0^{2j}\quad (j = 1, 2, \ldots). \tag{4}$$

Substituting the values of the coefficients obtained into expression (2), we get *Stirling's interpolating polynomial* which we designate as S_{2k}:

$$S_{2k}(t) = f_0 + \mu f_0^1 t + \frac{1}{2!} f_0^2 t^2 + \ldots + \frac{\mu f_0^{2k-1}}{(2k-1)!} t(t^2-1^2) \ldots (t^2-(k-1)^2) + \frac{f_0^{2k}}{(2k)!} t^2(t^2-1^2) \ldots (t^2-(k-1)^2). \tag{5}$$

Since Stirling's polynomial is only a new form of Lagrange's interpolating polynomial constructed with the use of the nodes $x_{-k}, \ldots, x_k$, it follows, by virtue of formula (4) from 7.5, that the remainder relative to the variable t can be represented as

$$R_{2k} = \frac{f^{(2k+1)}(\xi)}{(2k+1)!} h^{2k+1} \prod_{i=-k}^{k} (t-i),\quad \xi \in (x_{-k}, x_k), \tag{6}$$

and the estimate of the error of the approximate value $P_{2k}(x) = S_{2k}(t)$ (the error of the method) as

$$\Delta_1 = |f(x) - S(t)| \leqslant \frac{M_{2k+1}}{(2k+1)!} h^{2k+1} |t(t^2-1^2) \ldots (t^2-k^2)|, \tag{7}$$

where $M_{2k+1} = \max\limits_{[x_{-k},\, x_k]} |f^{(2k+1)}(x)|$.

Let now the interpolation point x^* lie between the nodes x_0 and x_1 close to the point $(x_0 + x_1)/2$. We have to construct an interpolating polynomial of an odd degree. Then, as we pointed out above, the net, minimizing the error, is symmetric with respect to the point $(x_0 + x_1)/2$, i.e. with respect to the point $t = 1/2$.

Thus, on the net

$$\ldots, x_{-k}, \ldots, x_{-1}, x_0, x_1, x_2, \ldots, x_{k+1}, \ldots$$

relative to the variable t, defined by formula (1), we construct an interpolating polynomial in the form

$$\begin{aligned} P_{2k+1}(x) &= P_{2k+1}(x_0 + th) = b_0 + \frac{b_1}{1!}\left(t - \frac{1}{2}\right) \\ &+ \frac{b_2}{2!}t(t-1) + \ldots + \frac{b_{2k}}{(2k)!}t(t^2 - 1^2) \\ &\cdot (t^2 - (k-1)^2)(t-k) \\ &+ \frac{b_{2k+1}}{(2k+1)!}t(t^2 - 1^2)\ldots(t^2 - (k-1)^2)\left(t - \frac{1}{2}\right)(t-k). \quad (8) \end{aligned}$$

The unknown coefficients b_i can be found from the conditions of coincidence of the values of the polynomial P_{2k+1} and the values of the function f at the nodes x_i.

Thus, to find the coefficients b_i, we have a system of linear equations

$$P_{2k+1}(x_0 + ih) = f_i \ (i = -k, \ldots, 0, 1, \ldots, k+1). \quad (9)$$

The structure of this system is such that its solution reduces to the successive solution of a system of two equations in two unknowns:

$$P_{2k+1}(x_0) = b_0 - \frac{b_1}{2} = f_0,$$

$$P_{2k+1}(x_1) = b_0 + \frac{b_1}{2} = f_1.$$

Hence $b_0 = \mu f_{1/2}$, $b_1 = f^1_{1/2}$.

Continuing this process, we note that at the jth step the determinant of the system of equations relative to the coefficients b_{2j-2} and b_{2j-1} is nonzero. It is easy to show this by analogy with what we did in constructing Stirling's polynomial. Consequently, all the coefficients b_i of polynomial (8) are uniquely defined by the system of equations (9) and, by virtue of the theorem on uniqueness, polynomial (8) is interpolating for the function f.

Simple transformations yield the following expressions for the coefficients:

$$b_{2j-2} = \mu f_{1/2}^{2j-2}, \quad b_{2j-1} = f_{1/2}^{2j-1} \quad (j = 1, 2, \ldots). \qquad (10)$$

Substituting the values of the coefficients we have found into expression (8), we get *Bessel's interpolating polynomial* which we designate as B_{2k+1}:

$$\begin{aligned} B_{2k+1}(t) = \mu f_{1/2} &+ \frac{f_{1/2}^{1}}{1!}\left(t-\frac{1}{2}\right) + \frac{\mu f_{1/2}^{2}}{2!}\, t\,(t-1) \\ &+ \frac{f_{1/2}^{3}}{3!}\, t\,(t-1)\left(t-\frac{1}{2}\right) + \\ \ldots &+ \frac{\mu f_{1/2}^{2k}}{(2k)!}\, t\,(t^2-1^2)\ldots(t^2-(k-1)^2)(t-k) \\ &+ \frac{f_{1/2}^{2k+1}}{(2k+1)!}\, t\,(t^2-1^2) \\ \ldots &(t^2-(k-1)^2)(t-k)\left(t-\frac{1}{2}\right). \end{aligned} \qquad (11)$$

Since Bessel's polynomial is another form of representation of Lagrange's interpolating polynomial constructed with the use of the nodes $x_{-k}, \ldots, x_{k+1}$, it follows, according to formula (4) from 7.5, that the remainder relative to the variable t can be written as

$$R_{2k+1} = \frac{f^{(2k+2)}(\xi)}{(2k+2)!}\, h^{2k+2} \prod_{i=-k}^{k+1} (t-i),$$

$$\xi \in (x_{-k}, x_{k+1}) \qquad (12)$$

and the estimate of the error of the approximate value $P_{2k+1}(x) = B_{2k+1}(t)$ (the error of the method) as

$$\Delta_1 = |\, f(x) - B_{2k+1}(t)\,| \leqslant \frac{M_{2k+2}}{(2k+2)!}\, h^{2k+2} \prod_{i=-k}^{k+1} |\, t-i\,|, \qquad (13)$$

where $M_{2k+2} = \max\limits_{[x_{-k},\, x_{k+1}]} |\, f^{(2k+2)}(x)\,|$.

We have thus considered two interpolating polynomials: Stirling's polynomial which is used to construct a polynomial of an even degree and is constructed with the use of an odd number of nodes and Bessel's polynomial which is used to construct a polynomial of an odd degree and is constructed with the use of an even number of nodes.

Now if the degree of a polynomial is not rigidly fixed, i.e. may be even as well as odd, then it is expedient to use Stirling's polynomial when

$$|t^*| = |x^* - x_0|/h \leqslant 0.25, \qquad (14)$$

i.e. when the interpolation point x^* lies closer to the node x_0 than to the middle point between the nodes. Bessel's polynomial should be used when

$$0.25 \leqslant t^* \leqslant 0.75, \qquad (15)$$

i.e. when the interpolation point x^* lies closer to the middle point between the nodes x_0 and x_1. One of the conditions (14) and (15) can also be ensured by the choice of the appropriate node as x_0.

Example. Using an appropriate interpolating polynomial, calculate, at the point $x_1^* = 48.63°$ and $x_2^* = 49.19°$ the values of the function $f = \sin x$ given as a table with a step of 1° (see Table 7.9), which contains the value f_i with four valid, in the narrow sense, digits. Estimate the error of the result.

△ We have established in Example 2 in 7.7 that the order of correctness of the table is 2. It is not advisable, therefore, to construct an interpolating polynomial of a degree higher than the second, i.e. it is expedient to construct either Stirling's polynomial of the second degree or Bessel's polynomial of the first degree.

Since the point $x_1^* = 48.63$ lies closer to the middle point between the nodes of 48° and 49°, we must take the node of 48° as x_0 when calculating $\sin x_1^*$, and make use of Bessel's polynomial. Then $t_1^* = (x_1^* - x_0)\,h^{-1} = 0.63$. The point $x_2^* = 49.19°$ lies close to the node of 49° and, therefore, we must take the nodal point $x_0 = 49°$ as the central node when calculating $\sin x_2^*$ and make use of Stirling's polynomial. Then $t_2^* = (x_2^* - x_0)\,h^{-1} = 0.19$.

Thus, employing our previous statements, from formulas (11) and (5) and also from the data given in Table 7.9, we have

$$B_1(0.63) = \frac{0.7431 + 0.7547}{2} + \frac{0.0116}{1!}\cdot 0.13 = 0.750408,$$

$$S_2(0.19) = 0.7547 + \frac{0.0116 + 0.0113}{2\cdot 1!}\cdot 0.19$$

$$+ \frac{-0.0003}{2!}\cdot 0.19^2 = 0.7568809.$$

We shall estimate now the errors of the method using formulas (13) and (7) respectively:

$$\Delta_1(B_1) < \frac{0.76}{2!}\left(\frac{\pi}{180}\right)^2 \cdot 0.63 \cdot 0.37 \approx 0.27 \cdot 10^{-4},$$

$$\Delta_1(S_2) < \frac{1}{3!}\left(\frac{\pi}{180}\right)^3 \cdot 0.19\,|0.19^2 - 1^2| \approx 0.2 \cdot 10^{-6}.$$

Taking into account the errors of the values of the function and of its finite differences given in Table 7.9, we get the values of the computational errors:

$$\Delta_2(B_1) = 0.5 \cdot 10^{-4} + 1 \cdot 10^{-4} \cdot 0.13 = 0.63 \cdot 10^{-4},$$

$$\Delta_2(S_2) = 0.5 \cdot 10^{-4} + 1 \cdot 10^{-4} \cdot 0.19 + 1 \cdot 10^{-4} \cdot 0.19^2 = 0.73 \cdot 10^{-4}.$$

When rounding off the values B_1 (0.63) and S_2 (0.19) to four decimal digits, we get the following rounding errors: $\Delta_3(B_1) = 0.08 \cdot 10^{-4}$ and $\Delta_3(S_2) = 0.11 \cdot 10^{-4}$.

Combining all the errors we have found, we obtain $\sin 48.63° \approx 0.7509 \pm 0.0001$, $\sin 49.19° = 0.7569 \pm 0.0001$. ▲

Remark. In some cases, to minimize the total error, it is expedient to use an interpolating polynomial whose degree is higher than the order of correctness of the table of finite differences. This can be explained as follows. Naturally, the computing error increases with an increase in the degree of a polynomial. At the same time, the decrease in the error of the method can be so sharp that it involves a decrease in the total error.

7.9. Newton's First and Second Interpolating Polynomials

If the interpolation point x^* is at the beginning or the end of the table, then it is not always possible to choose a sufficient number of nodes to the left and to the right of x^* to construct the necessary finite differences. In that case we use special forms of an interpolating polynomial.

Let the point x^* lie close to the first node of the net $x_0, x_1, \ldots, x_k, \ldots$. We consider a variable t defined by relation (1) from 7.8 and construct an interpolating polynomial in the following form:

$$P_k(x) \doteq P_k(x_0 + th) = a_0 + \frac{a_1}{1!}t + \frac{a_2}{2!}t(t-1) + \ldots + \frac{a_k}{k!}t\ldots(t-k+1). \quad (1)$$

We find the unknown coefficients a_i from the conditions of coincidence of the polynomial P_k and the function f at the nodes x_i. Recall that each node x_i is associated with the quantity $t = i$. Thus, to find the coefficients a_i, we get a system of linear equations

$$P_k(x_0 + ih) = f_i \ (i = 0, 1, \ldots, k). \tag{2}$$

The structure of this system is such that a_0 can be found directly from the first equation of system (2), a_1 can be found from the second equation for a_0 already found and so on. Indeed, setting $i = 0$, we find, from the first equation of system (2), that

$$P_k(x_0) = a_0 = f_0,$$

from the second equation we find, for $i = 1$, that

$$P_k(x_0 + h) = f_0 + \frac{a_1}{1!} \cdot 1 = f_1, \quad a_1 = \Delta f_0.$$

Continuing this process, we get, as a result of simple transformations, the following expressions for the coefficients:

$$a_0 = f_0, \ a_i = \Delta^i f_0 \ (i = 1, 2, \ldots, k). \tag{3}$$

Substituting the values of the coefficients we have found into relation (1), we get *Newton's first interpolating polynomial* which we designate as N_k^{I}:

$$N_k^{\mathrm{I}}(t) = f_0 + \frac{\Delta f_0}{1!} t + \frac{\Delta^2 f_0}{2!} t(t-1) + \ldots + \frac{\Delta^k f_0}{k!} t(t-1) \ldots (t-k+1). \tag{4}$$

By virtue of formula (4) from 7.5, we can represent the remainder relative to the variable t in the form

$$R_k = \frac{f^{(k+1)}(\xi)}{(k+1)!} h^{k+1} t(t-1) \ldots (t-k), \ \xi \in (x_0, x_k) \tag{5}$$

and the estimate of the error of the approximation $N_k^{\mathrm{I}}(t)$ (the error of the method) in the form

$$\Delta_1 = | f(x) - N_k^{\mathrm{I}}(t) | \leqslant \frac{M_{k+1}}{(k+1)!} h^{k+1} | t(t-1) \ldots (t-k) |, \tag{6}$$

where $M_{k+1} = \max\limits_{[x_0, x_k]} | f^{(k+1)}(x) |$.

Let now the point x^* lie close to the last node of the net $\ldots x_{-k}, \ldots, x_{-1}, x_0$. Again using the variable t, defined by relation (1) from 7.8, we construct an interpolating polynomial for this net in the form

$$P_k(x) = P_k(x_0 + th)$$
$$= a_0 + \frac{a_1}{1!} t + \frac{a_2}{2!} t(t+1) + \ldots + \frac{a_k}{k!} t \ldots (t+k-1). \tag{7}$$

The unknown coefficients a_i can be found from the conditions of coincidence of the polynomial P_k and the function f at the nodes x_i. Note that the node x_{-i} is associated with the value $t = -i$. Thus we get a system of linear equations

$$P_k(x_0 - ih) = f_i \quad (i = 0, 1, \ldots, k) \tag{8}$$

to obtain the coefficients a_i. The structure of this system is such that a_0 can be found directly from the first equation of system (8), a_1 from the second equation for a_0 already found and so on. Indeed, setting $i = 0$, we find from the first equation of system (8) that

$$P_k(x_0) = a_0 = f_0$$

and from the second equation, for $i = 1$, we have

$$P_k(x_0 - h) = f_0 + \frac{a_1}{1!} 1 = f_{-1}, \quad a_1 = \nabla f_0.$$

Continuing this process, we get, as a result of simple transformations, the following expressions for the coefficients:

$$a_0 = f_0, \quad a_i = \nabla^i f_0 \quad (i = 1, 2, \ldots, k). \tag{9}$$

Substituting the values of the coefficients we have found into relation (7), we get *Newton's second interpolating polynomial* which we designate as N_k^{II}:

$$N_k^{II}(t) = f_0 + \frac{\nabla f_0}{1!} t + \frac{\nabla^2 f_0}{2!} t(t+1) +$$
$$\ldots + \frac{\nabla^k f_0}{k!} t(t+1) \ldots (t+k-1) \tag{10}$$

with the remainder

$$R_k = \frac{f^{(k+1)}(\xi)}{(k+1)!} h^{k+1} t(t+1)\dots(t+k), \qquad (11)$$
$$\xi \in (x_{-k}, x_0),$$

and the estimate of the approximation error

$$\Delta_1 = |f(x) - N^{II}(t)| \leqslant \frac{M_{k+1}}{(k+1)!} h^{k+1} |t(t+1)\dots(t+k)|, \qquad (12)$$

where $M_{k+1} = \max\limits_{[x_{-k}, x_0]} |f^{(k+1)}(x)|$.

Formulas (4) and (10) are often called *Newton's interpolation formulas for the forward and backward interpolation* respectively.

Example. Set up appropriate interpolating polynomials and calculate, at the points $x_1^* = 0.63$ and $x_2^* = 1.35$, the values of the function $f = 3^x$ given as a table which contains the values f_i with four valid, in the broad sense, digits:

x	0.50	0.75	1.00	1.25	1.50
f	1.732	2.280	3.000	3.948	5.196

Evaluate the error of the result.

△ We put in this table the values of finite differences writing the values of the errors of the corresponding finite differences in the last row:

Table 7.10

x	f	f^1	f^2	f^3	f^4
0.50	1.732				
		548			
0.75	2.280		172		
		720		56	
1.00	3.000		228		16
		948		72	
1.25	3.948		300		
		1248			
1.50	5.196				
	0.001	2	4	8	16

Since the fourth finite difference coincides with its error, it is not expedient from the point of view of the calculation error, to approximate the given function by a polynomial of a degree higher than the third.

Furthermore, since $x_1^* = 0.63$ is at the beginning of the table and $x_2^* = 1.35$ is at its end, it follows that we must use Newton's first interpolating polynomial to calculate $f_1^* = 3^{0.63}$ and Newton's second interpolating polynomial to calculate $f_2^* = 3^{1.35}$.

Thus, setting $x_0 = 0.5$, we calculate $t_1^* = (x_1^* - x_0)/h = (0.63-0.5)/0.25 = 0.52$. Substituting the value of t_1^* obtained into expression (4) for Newton's first interpolating polynomial and using the values of finite differences given in Table 7.10, we obtain

$$N_3^{\mathrm{I}}(0.52) = 1.732 + \frac{0.548}{1!}\cdot 0.52 + \frac{0.172}{2!}\cdot 0.52\cdot(-0.48)$$

$$+\frac{0.056}{2}\cdot 0.52\,(-0.48)\,(-1.48) = 1.9989420.$$

Similarly, setting $x_0 = 1.50$, we calculate $t_2^* = (1.35-1.50)/0.25 = -0.60$ and, using expression (10) for Newton's second interpolating polynomial, we obtain

$$N_3^{\mathrm{II}}(-0.60) = 5.196 + \frac{1.248}{1!}\cdot(-0.60) + \frac{0.300}{2!}\cdot(-0.60)\cdot 0.40$$

$$+\frac{0.072}{3!}\cdot(-0.60)\cdot 0.40\cdot 1.40 = 4.407168.$$

Let us use formulas (6) and (12) to estimate the error of the method:

$$\Delta_1(N_3^{\mathrm{I}}) < \frac{4\ln^4 3}{4!}\cdot 0.25^4\cdot 0.52\cdot 0.48\cdot 1.48\cdot 2.48 = 0.0009,$$

$$\Delta_1(N_3^{\mathrm{II}}) < \frac{5.2\ln^4 3}{4!}\cdot 0.25^4\cdot 0.60\cdot 0.40\cdot 1.40\cdot 2.40 = 0.001.$$

Taking into account the values of the errors f^k given in Table 7.10, we evaluate the computational errors:

$$\Delta_2(N_3^{\mathrm{I}}) < 0.001 + 0.0011 + 0.0005 + 0.0005 = 0.0031,$$

$$\Delta_2(N_3^{\mathrm{II}}) < 0.001 + 0.0012 + 0.0005 + 0.0005 = 0.0032.$$

Rounding off the values $N_3^{\mathrm{I}}(0.52)$ and $N_3^{\mathrm{II}}(-0.60)$ to four decimal digits, we get the following rounding errors:

$$\Delta_3(N_3^{\mathrm{I}}) = 0.06\cdot 10^{-3}, \qquad \Delta_3(N_3^{\mathrm{II}}) = 0.2\cdot 10^{-3}.$$

Combining all the errors found, we finally have $3^{0.63} \approx 1.999 \pm 0.005$, $3^{1.35} \approx 4.407 \pm 0.005$. ▲

7.10. Divided Differences

In the preceding sections we considered various forms of interpolating polynomial for a uniform net of nodes. The coefficients of the polynomials constructed were found with the aid of finite differences. In this section we again consider the case when the values of a function are given at nonequispaced nodes. In such a case, instead of finite differences we consider *divided differences* which are, in a sense, an analogue of the concept of a derivative and are defined as follows.

Assume that the function $y = f(x)$ is defined by its values $y_0 = f_0 = f(x_0)$, $y_1 = f_1 = f(x_1)$, . . ., $y_k = f_k = f(x_k)$, . . . at the nodes x_i of an arbitrary net Λ_n. The *divided differences of order zero* $f(x_i)$ coincide with the values of the function at the points x_i. Relations of the form

$$f(x_0, x_1) = \frac{f_1 - f_0}{x_1 - x_0}, \quad f(x_1, x_2) = \frac{f_2 - f_1}{x_2 - x_1},$$
$$\ldots, \quad f(x_i, x_{i+1}) = \frac{f_{i+1} - f_i}{x_{i+1} - x_i}, \quad \ldots$$

are known as the *first divided differences*. Relations of the form

$$f(x_0, x_1, x_2) = \frac{f(x_1, x_2) - f(x_0, x_1)}{x_2 - x_0},$$
$$f(x_1, x_2, x_3) = \frac{f(x_2, x_3) - f(x_1, x_2)}{x_3 - x_1},$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$f(x_i, x_{i+1}, x_{i+2}) = \frac{f(x_{i+1}, x_{i+2}) - f(x_i, x_{i+1})}{x_{i+2} - x_i}$$

are the *second divided differences*. In the general case, the *kth divided difference* can be found with the use of the $(k-1)$th divided difference from the formula

$$f(x_i, x_{i+1}, \ldots, x_{i+k}) = \frac{f(x_{i+1}, \ldots, x_{i+k}) - f(x_i, \ldots, x_{i+k-1})}{x_{i+k} - x_i}. \tag{1}$$

Other designations can be used for divided differences (difference quotients):

$$f(x_i, x_{i+1}, \ldots, x_{i+k}) \equiv [x_i, x_{i+1}, \ldots, x_{i+k}].$$

It is convenient to tabulate divided differences by analogy with what was done for central finite differences:

Table 7.11

x_i	$y_i=f(x_i)$	$[x_i, x_{i+1}]$	$[x_i, x_{i+1}, x_{i+2}]$	$[x_i, x_{i+1}, x_{i+2}, x_{i+3}]$	$[x_i, x_{i+1}, x_{i+2}, x_{i+3}, x_{i+4}]$
x_0	y_0				
		$[x_0, x_1]$			
x_1	y_1		$[x_0, x_1, x_2]$		
		$[x_1, x_2]$		$[x_0, x_1, x_2, x_3]$	
x_2	y_2		$[x_1, x_2, x_3]$		$[x_0, x_1, x_2, x_3, x_4]$
		$[x_2, x_3]$		$[x_1, x_2, x_3, x_4]$	
x_3	y_3		$[x_2, x_3, x_4]$		
		$[x_3, x_4]$			
x_4	y_4				

Example 1. Compile a table of divided differences for the function $y=f(x)$ given as the following table:

x	0	1	5	10
y	10	20	100	1100

△ Using the definition, we find the first divided differences

$$[x_0, x_1]=\frac{y_1-y_0}{x_1-x_0}=\frac{20-10}{1-0}=10,$$

$$[x_1, x_2]=\frac{y_2-y_1}{x_2-x_1}=\frac{100-20}{5-1}=\frac{80}{4}=20,$$

$$[x_2, x_3]=\frac{y_3-y_2}{x_3-x_2}=\frac{1100-100}{10-5}=\frac{1000}{5}=200.$$

In a similar way we find the second divided differences:

$$[x_0, x_1, x_2]=\frac{[x_1, x_2]-[x_0, x_1]}{x_2-x_0}=\frac{20-10}{5-0}=2.$$

$$[x_1, x_2, x_3]=\frac{[x_2, x_3]-[x_1, x_2]}{x_3-x_1}=\frac{200-20}{10-1}=20.$$

The third divided difference

$$[x_0, x_1, x_2, x_3]=\frac{[x_1, x_2, x_3]-[x_0, x_1, x_2]}{x_3-x_0}=\frac{20-2}{10}=1.8.$$

We tabulate the results of calculations:

x	y	$[x_i, x_{i+1}]$	$[x_i, x_{i+1}, x_{i+2}]$	$[x_i, x_{i+1}, x_{i+2}, x_{i+3}]$
0	10			
		10		
1	20		2	
		20		1.8
5	100		20	
		200		
10	1100			

Here are some properties of divided differences.

1°. *Divided differences of all orders are linear combinations of the values $f_i = f(x_i)$, namely, the following formula holds true:*

$$f(x_0, \ldots, x_k)=\sum_{i=0}^{k}\frac{f_i}{\prod\limits_{\substack{j=0\\ j\neq i}}^{k}(x_i-x_j)}. \tag{2}$$

☐ For $k=0$ we get an identity $f(x_0)=f_0$, for $k=1$ we have $f(x_0, x_1)=\frac{f(x_0)-f(x_1)}{x_0-x_1}$. And this is the definition of the divided difference $f(x_0, x_1)$.

We shall continue with the proof by induction. Assume that equality (2) is valid for all $k\leqslant n$. Then

$$f(x_0, \ldots, x_{n+1})=\frac{f(x_0, \ldots, x_n)-f(x_1, \ldots, x_{n+1})}{x_0-x_{n+1}}$$

$$=\frac{1}{x_0-x_{n+1}}\left(\sum_{i=0}^{n}\frac{f_i}{\prod\limits_{\substack{j=0\\ j\neq i}}^{n}(x_i-x_j)}-\sum_{i=1}^{n+1}\frac{f_i}{\prod\limits_{\substack{j=1\\ j\neq i}}^{n+1}(x_i-x_j)}\right)$$

$$= \frac{f_0}{\prod\limits_{j=1}^{n+1} (x_0 - x_j)} + \sum_{i=1}^{n} \frac{f_i}{\prod\limits_{\substack{j=0\\ j\neq i}}^{n+1} (x_i - x_j)} + \frac{f_{n+1}}{\prod\limits_{j=0}^{n} (x_{n+1} - x_j)}$$

$$= \sum_{i=0}^{n+1} \frac{f_i}{\prod\limits_{\substack{j=0\\ j\neq i}}^{n+1} (x_i - x_j)}.$$

Thus equality (2) holds true for $k = n + 1$ as well. ■

2°. *A divided difference is a symmetric function of its arguments, i.e. it does not change upon their permutation.*

□Upon any permutation of the arguments $x_0, \ldots, x_k$, the corresponding terms on the right-hand side of relation (2) only change places but the sum evidently remains unchanged. Consequently, the left-hand side of relation (2) does not change either, i.e. $f(x_0, \ldots, x_k)$. ■

3°. *A divided difference satisfies the equality*

$$(af + bg)(x_0, \ldots, x_k) = af(x_0, \ldots, x_k) + bf(x_0, \ldots, x_k), \qquad (3)$$

where a and b are constants.

This immediately follows from relation (2) since its right-hand side is linear with respect to f_i.

The next property expresses the connection between the divided difference $f(x_0, \ldots, x_k)$ and the derivative $f^k(x)$.

4°. *If the nodes $x_0, \ldots, x_k$ belong to the interval $[a, b]$ and the function $f(x)$ has a continuous kth-order derivative on $[a, b]$, then there is a point $\xi \in (a, b)$ such that*

$$f(x_0, \ldots, x_k) = \frac{1}{k!} f^{(k)}(\xi). \qquad (4)$$

This property yields a simple corollary. Let $f(x) = a_0x^k + a_1x^{k-1} + \ldots + a_k$ be a polynomial of degree k. Then, evidently, $f^{(k)}(x) = a_0k!$ and relation (4) yields the following value for the divided difference:

$$f(x_0, \ldots, x_k) = \frac{1}{k!} a_0k! = a_0. \qquad (5)$$

Thus for any polynomial of degree k the kth-order divided differences are equal to a constant quantity, which is the coefficient in the leading power of the polynomial. Divided differences of higher orders (higher than k) are evidently zero.

Example 2. Let us verify the validity of Property 1° using four values of the function $y = f(x)$, i.e. $y_i = f(x_i)$ $(i = 0, 1, 2, 3)$.

We find the first divided differences:

$$f(x_0, x_1) = \frac{y_0 - y_1}{x_0 - y_1}, \quad f(x_1, x_2) = \frac{y_1 - y_2}{x_1 - x_2}, \quad f(x_2, x_3) = \frac{y_2 - y_3}{x_2 - x_3},$$

and then seek the second divided differences:

$$f(x_0, x_1, x_2) = \frac{1}{x_0 - x_2}\left(\frac{y_0 - y_1}{x_0 - x_1} - \frac{y_1 - y_2}{x_1 - x_2}\right)$$

$$= \frac{y_0}{(x_0 - x_1)(x_0 - x_2)} + \frac{y_1}{(x_1 - x_0)(x_1 - x_2)} + \frac{y_2}{(x_2 - x_0)(x_2 - x_1)},$$

$$f(x_1, x_2, x_3) = \frac{1}{x_1 - x_3}\left(\frac{y_1 - y_2}{x_1 - x_2} - \frac{y_2 - y_3}{x_2 - x_3}\right)$$

$$= \frac{y_1}{(x_1 - x_2)(x_1 - x_3)} + \frac{y_2}{(x_2 - x_1)(x_2 - x_3)} + \frac{y_3}{(x_3 - x_1)(x_3 - x_2)}.$$

This is in full accord with Property 1°. We can show in a similar way that the third divided difference

$$f(x_0, x_1, x_2, x_3) = \frac{y_0}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)} + \frac{y_1}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)} + \frac{y_2}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)} + \frac{y_3}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)}.$$

We calculate the third finite difference using the values of the function and the argument given in Example 1:

$$[x_0, x_1, x_2, x_3] = \frac{10}{(-1)(-5)(-10)} + \frac{20}{1\cdot(-4)(-9)} + \frac{100}{5\cdot 4\cdot(-5)} + \frac{1100}{10\cdot 9\cdot 5} = -\frac{1}{5} + \frac{5}{9} - 1 + \frac{22}{9} = 1.8.$$

This coincides with the value obtained in Example 1.

Example 3. Let us verify now the validity of Property 2° for the function $y = f(x)$. We shall show, for instance, that $f(x_0, x_1, x_2) = f(x_1, x_0, x_2)$.

Indeed, according to the definition of a divided difference we have

$$f(x_1, x_0, x_2) = \frac{1}{x_1 - x_2}\left(\frac{y_1 - y_0}{x_1 - x_0} - \frac{y_0 - y_2}{x_0 - x_2}\right)$$
$$= \frac{y_1}{(x_1 - x_0)(x_1 - x_2)} + \frac{y_0}{(x_0 - x_1)(x_0 - x_2)}$$
$$+ \frac{y_2}{(x_2 - x_0)(x_2 - x_1)}.$$

We have obtained the same result as in Example 2 (with a change in the sequence of terms).

Example 4. We shall illustrate Property 3°. Together with the function $y = f(x)$ given in Example 2 we shall consider a function $z = g(x)$ and its values $z_i = g(x_i)$ $(i = 0, 1, 2, 3)$ specified at the same nodes as y_i. We set up a linear combination $u(x) = af(x) + bg(x)$, where a and b are constants. Let us now calculate $u(x_0, x_1, x_2)$. As in Example 2 we find that

$$u(x_0, x_1, x_2) = \frac{ay_0 + bz_0}{(x_0 - x_1)(x_0 - x_2)} + \frac{ay_1 + bz_1}{(x_1 - x_0)(x_1 - x_2)}$$
$$+ \frac{ay_2 + bz_2}{(x_2 - x_0)(x_2 - x_1)}.$$

Grouping separately the first and the second summands of each of the three terms of the sum, we obtain

$$u(x_0, x_1, x_2) = af(x_0, x_1, x_2) + bg(x_0, x_1, x_2),$$

and this is what we wished to prove.

7.11. Newton's Interpolating Polynomial for an Arbitrary Net of Nodes

Using Lagrange's form, we represent the interpolating polynomial as

$$L_0(x) + (L_1(x) - L_0(x)) + \dots + (L_n(x) - L_{n-1}(x).$$

Here $L_0(x) = f(x_0)$, $L_k(x)$ $(k = 1, 2, \dots, n)$ are interpolating polynomials in Lagrange's form constructed with the use of the nodes $x_0, x_1, \dots, x_k$. We consider the differences

$$L_k - L_{k-1} = \sum_{i=0}^{k} f_i \frac{\omega_k(x)}{(x - x_i)\,\omega_k'(x_i)} - \sum_{i=0}^{k-1} f_i \frac{\omega_{k-1}(x)}{(x - x_i)\,\omega_{k-1}'(x_i)}$$
$$= \sum_{i=0}^{k-1} f_i \frac{\omega_{k-1}(x)}{\omega_k'(x_i)} + f_k \frac{\omega_{k-1}(x)}{\omega_k'(x_k)} = \omega_{k-1}(x) \sum_{i=0}^{k} \frac{f_i}{\omega_k'(x_i)}.$$

Thus, using formula (2) from 7.10, we obtain

$$L_k - L_{k-1} = \omega_{k-1}(x)\, f(x_0, \ldots, x_k) \qquad (1)$$

and the interpolating polynomial assumes the form

$$N_n(x) = f_0 + (x - x_0)\, f(x_0, x_1) + \ldots + (x - x_0) \ldots (x - x_{n-1})\, f(x_0, \ldots, x_n). \qquad (2)$$

This form is known as *Newton's interpolating polynomial with divided differences.*

The expression for the error has evidently the same form as in the case of Lagrange's polynomial [see formulas (8) and (9) from 7.5]

Example 1. Using the nodes $x_0 = 0$, $x_1 = 1/3$, $x_2 = 1$, construct Newton's interpolating polynomial for the function $f = \sin(\pi x/2)$.

△ Taking into account that $f_0 = 0$, $f_1 = 0.5$, $f_2 = 1$, we set up the necessary divided differences:

$$f(x_0, x_1) = \frac{0 - \frac{1}{2}}{0 - \frac{1}{3}} = \frac{3}{2}, \quad f(x_1, x_2) = \frac{\frac{1}{2} - 1}{\frac{1}{3} - 1} = \frac{3}{4},$$

$$f(x_0, x_1, x_2) = \frac{\frac{3}{2} - \frac{3}{4}}{0 - 1} = -\frac{3}{4}.$$

Substituting the values obtained into formula (2), we have, for $n = 2$, the following polynomial:

$$N_2(x) = 0 + \frac{3}{2}(x - 0) - \frac{3}{4}(x - 0)\left(x - \frac{1}{3}\right).$$

This polynomial must evidently be identical to Lagrange's polynomial constructed in Example 1 in 7.6. The reader is invited to prove this independently. ▲

Note that in form (2) of interpolating polynomial a unique condition is imposed on the nodes x_i, their noncoincidence. Therefore, the nodes may be enumerated arbitrarily. For instance, the index "0" often denotes the last node in the table, x_{-1}, its last but one node and so on. In this case polynomial (2) assumes the form

$$N_n(x) = f_0 + (x - x_0) f(x_0, x_{-1}) + \ldots + (x - x_0) \ldots (x - x_{-n+1}) f(x_0, \ldots, x_{-n}) \qquad (3)$$

and is called *Newton's polynomial for backward inerpolation.*

To illustrate what we have stated, we shall solve Example 1 for $x_0 = 1$, $x_{-1} = 1/3$, $x_{-2} = 0$. From formula (3) we obtain

$$N_2(x) = 1 + \frac{3}{4}(x-1) - \frac{3}{4}(x-1)\left(x - \frac{1}{3}\right).$$

Note that the necessary divided differences have been taken from Example 1, the property of their symmetry being used with respect to their arguments.

The comparison of Lagrange's and Newton's forms for an interpolating polynomial makes it possible to recommend the use of representation in Lagrange's form, first, in theoretical research, say, to study the problem of convergence of $P_n(f, x)$ to f as $n \to \infty$, second, to interpolate several functions on the same net of nodal points since in this case Lagrange's multipliers l_i can be calculated once and then used to interpolate all the functions.

Representation in Newton's form proves to be the most convenient in practical computations. Indeed, the number of nodal points to be used and the degree of the interpolating polynomial are often not predetermined and when we pass from n nodes to $n + 1$ nodes in Newton's form we add only one term which has the sense of a correction of the value already calculated, whereas in Lagrange's form an addition of one more term must be followed by a complete recalculation of the result obtained. In addition, in computations interpolation is usually carried out on a small interval of length $h < 1$ and the summands in Newton's form are of the order of $h^0, h^1, h^2, \ldots$, i.e. are arranged in a decreasing order, which is convenient when the accuracy of the result of interpolation is determined.

7.12. Practical Interpolation in Tables

Linear or quadratic interpolation is usually used when interpolation is carried out in tables.

In the case of linear interpolation the value of a function at a point different from the interpolation nodes is found from two known values of the tabulated function $y_i = f(x_i)$, $y_{i+1} = f(x_{i+1})$ at the interpolation nodes x_i

and x_{i+1} between which the value of the required argument x lies.

Lagrange's interpolation formula for linear interpolation assumes the form

$$L_1(x) = y_i \frac{x - x_{i+1}}{x_i - x_{i+1}} + y_{i+1} \frac{x - x_i}{x_{i+1} - x_i},$$

and Newton's first interpolation formula assumes the form

$$N_1(x) = y_i + \frac{\Delta y_i}{h}(x - x_i),$$

where $\Delta y_i = y_{i+1} - y_i$ is the first finite difference at the point x_i and $h = x_{i+1} - x_i$ is the stepsize of interpolation.

Thus, to use Newton's formula to obtain an approximate value of the function $y(x)$, it is sufficient to add a correction equal to $\Delta y_i (x - x_i) h^{-1}$ to the tabulated value y_i.

Example 1. Find out how many degrees are there in the radian measure 0.222?

Δ We use the table

Radians	Degrees
0.22	12.605
23	13.178
24	13.751

To carry out a linear interpolation, it is sufficient to consider the data in the first two rows. We set up a tabular difference

$$\Delta y_i = y_{i+1} - y_i = 13.178 - 12.605 = 0.573.$$

The stepsize of the table $h = 0.01$, $x - x_i = 0.222 - 0.220 = 0.002$. We calculate the correction

$$\frac{\Delta y_i}{h}(x - x_i) = \frac{0.573}{0.01} \cdot 0.002 = 0.1146$$

and add it to the tabular value:

$$y = 12.605 + 0.1146 = 12.7196.$$

Let us verify the error of the value obtained. Using formula (8) from 7.5, we have the following value for the error of the method:

$$\Delta_1 = \frac{M_2}{2!} \,|(0.222 - 0.22)(0.222 - 0.23)|.$$

Since the function being interpolated is $y = (180\, x/\pi)$, it follows that $M_2 = 0$ and $\Delta_1 = 0$.

Let us find the computing error, taking into account that the error of the initial data constitutes 0.0005:

$$\Delta_2 = 0.0005 + \frac{0.001}{0.01} \cdot 0.002 = 0.0007.$$

Rounding off the value y (0.222) to three decimal digits, we get $\Delta_3 = 0.0004$.

Summing up all the errors found, we finally have $y(0.222) = 12.720 \pm 0.0011$. ▲

In the case of a quadratic interpolation we use three values of the tabulated function $y_{-1} = f(x_{-1})$, $y_0 = f(x_0)$, $y_1 = f(x_1)$. The interpolating polynomial is constructed either in the form of Lagrange (for nonequispaced nodes) or in Newton's form when the interpolation point is closer to x_{-1} or to x_1 than to x_0, or in Stirling's form when the interpolation point is close to the node x_0.

Example 2. Construct Newton's interpolating polynomial for the function $y = \ln x$ using its tabulated values

x	2	3	5
y	0.6931	1.0986	1.6094

and obtain a uniform estimate of the error on the interval [2, 5].

△ First of all we seek divided differences:

$$f(2,\ 3) = \frac{1.0986 - 0.6931}{3-2} = 0.4055,$$

$$f(3,\ 5) = \frac{1.6094 - 1.0986}{5-3} = 0.2554,$$

$$f(2,\ 3,\ 5) = \frac{0.2554 - 0.4055}{5-2} = -0.0500.$$

Substituting the values obtained into formula (2) from 7.11, we find, for $n = 2$, that

$$N_2(x) = 0.6931 + 0.4055\,(x-2) - 0.0500\,(x-2)(x-3).$$

Using estimate (8) from 7.5, we have for the error of the method

$$\Delta_1 = \frac{M_3}{3!} \max_{[2,\ 5]} |(x-2)(x-3)(x-5)|.$$

Next we find that

$$M_3 = \max_{[2,\ 5]} \left| \frac{2}{x^3} \right| = \frac{1}{4}, \quad \max_{[2,\ 5]} |(x-2)(x-3)(x-5)| \cong 2.2.$$

Thus $\Delta_1 = 0.1$.

The computing error is evidently negligibly small as compared to the error of the method. Therefore the maximum possible error of interpolation is 0.1. ▲

7.13. Aitken's Iterated Interpolation

It is expedient to employ Aitken's iterated interpolation when it is not necessary to obtain an approximate analytic expression for the function $f(x)$, given in a tabular form, but the only problem is to find the value of this function at a point x^* different from the interpolation nodes. The gist of the method is a successive linear interpolation. The process of computing $f(x^*)$ is the following. We enumerate all the interpolation nodes, say, in the order of their receding from x^* and compile the following table.

Table 7.12

.	.				
.	.				
.					
x_4	P_0^4	$P_1^{2.4}$	P_2^{024}	P_3^{0124}	
x_2	P_0^2	$P_1^{0.2}$	P_2^{012}	P_3^{0123}	...
x_0	P_0^0	$P_1^{1.0}$	P_2^{013}		
x_1	P_0^1	$P_1^{1.3}$			
x_3	P_0^3				
.	.				
.	.				
.	.				

Here

$$P_0^k = f(x_k),$$

$$P_1^{ij}(x) = f_i \frac{x-x_j}{x_i-x_j} + f_j \frac{x-x_i}{x_j-x_i} = \frac{1}{x_j-x_i} \begin{vmatrix} x-x_i & P_0^i \\ x-x_j & P_0^j \end{vmatrix}$$

is an interpolating polynomial of a degree not higher than the first, constructed with the use of the nodal points x_i and x_j,

$$P_2^{ijk}(x) = \frac{1}{x_k-x_i} \begin{vmatrix} x-x_i & P_1^{ij} \\ x-x_k & P_1^{jk} \end{vmatrix}$$

is an interpolating polynomial of a degree not higher than the second constructed with the use of the nodal points x_i, x_j, x_k. Continuing this process, we construct a polynomial

$$P_n^{ij\ldots km}(x) = \frac{1}{x_m - x_i}\begin{vmatrix} x - x_i & P_{n-1}^{ij\ldots k}(x) \\ x - x_m & P_{n-1}^{j\ldots km}(x) \end{vmatrix}. \qquad (1)$$

We shall show that if $P_{n-1}^{ij\ldots k}(x)$ and $P_{n-1}^{j\ldots km}(x)$ are interpolating polynomials constructed with the use of the nodal points x_i, $x_j, \ldots, x_k$ and $x_j, \ldots, x_k, x_m$ respectively, then $P_n^{ij\ldots km}(x)$ is an interpolating polynomial constructed with the use of the nodal points $x_i, x_j, \ldots, x_k, x_m$.

Indeed, first, $P_n^{ij\ldots km}(x)$ is a polynomial of a degree not higher than n. This can be seen from the construction of formula (1). Second, at all nodal points x_p the polynomial $P_n^{ij\ldots km}(x)$ assumes the corresponding value:

$$P_n^{ij\ldots km}(x_i) = \frac{-(x_i - x_m) f_i}{x_m - x_i} = f_i \; (x_p = x_i),$$

$$P_n^{ij\ldots}(x_m) = \frac{(x_m - x_i) f_m}{x_m - x_i} = f_m \; (x_p = x_m),$$

$$P_n^{ij\ldots km}(x_p) = \frac{1}{x_m - x_i}\left((x_p - x_i) f_p - (x_p - x_m) f_p\right) = f_p.$$

Calculating successively the values $P_n^{01\ldots n}(x^*)$ from formula (1), we take them to be the successive approximations of $f(x^*)$. The computing process is terminated when the absolute value of the difference of two successive approximations becomes sufficiently small.

Example 1. At the point $x^* = 6$ calculate, with an accuracy of $\varepsilon = 0.05$, the value of the function $f = \ln x$ given as a table

x	1	2	4	5	8	10
f	0.00	0.69	1.39	1.61	2.08	2.30

△ We enumerate the nodes as follows: $x_0 = 5$, $x_1 = 8$, $x_2 = 4$, $x_3 = 10$, $x_4 = 2$, $x_5 = 1$. Using formula (1), we calculate the values of the interpolating polynomials P_n (6):

$$P_1^{01} = \frac{1}{8-5}\cdot\begin{vmatrix} 6-5 & 1.61 \\ 6-8 & 2.08 \end{vmatrix} = 1.77,$$

$$P_1^{02} = \frac{1}{5-4}\cdot\begin{vmatrix} 6-4 & 1.39 \\ 6-5 & 1.61 \end{vmatrix} = 1.83,$$

$$P_2^{012} = \frac{1}{8-4}\cdot\begin{vmatrix} 6-4 & 1.83 \\ 6-8 & 1.77 \end{vmatrix} = 1.80.$$

Since $|P_1^{02} - P_2^{012}| = 0.03 < 0.05$, we terminate the calculations and set $\ln 6 = 1.80 \pm 0.03$. ▲

Example 2. Using Aitken's scheme, calculate, with an accuracy of $0.5\cdot 10^{-4}$, the value of sin 0.674 for the function $y = \sin x$, given as a table

$x_0 = 0.66$	$x_1 = 0.67$	$x_2 = 0.68$
$y_0 = 0.61312$	$y_1 = 0.62099$	$y_2 = 0.62879$

△ According to formula (1) we have

$$P_1^{01}(0.674) = \frac{\begin{vmatrix} 0.674-0.68 & 0.61312 \\ 0.674-0.67 & 0.62090 \end{vmatrix}}{0.67-0.66} = 0.625730,$$

$$P_1^{12} = \frac{1}{0.68-0.67}\cdot\begin{vmatrix} 0.674-0.67 & 0.62099 \\ 0.674-0.68 & 0.625643 \end{vmatrix} = 0.625643,$$

$$P_2^{012} = \frac{1}{0.68-0.66}\cdot\begin{vmatrix} 0.674-0.66 & 0.625730 \\ 0.674-0.68 & 0.625643 \end{vmatrix} = 0.625676.$$

Consequently, $\sin 0.674 = 0.62568 + 0.00004$. ▲

7.14. "Optimal-Interval" Interpolation

Let us consider again the estimate of the error expressed by formula (8) from 7.5. We assume now that there are no restrictions in the choice of the net Λ_n. We pose a problem of the best choice of interpolation nodes. Proceeding from the estimate (8) from 7.5, we find that the best interpolation nodes for the class of functions $C^{n+1}([a, b])$ being considered are x_i for which the expression $\max\limits_{[a,b]} |\omega_n(x)|$ is minimum. The determination of these nodes actually reduces to finding the roots of a polynomial which has the smallest deviation from zero on the interval $[a, b]$. As is known from the theory of functions, such a polynomial is generated by *Chebyshev's polynomials of the first kind* which are defined by the following recurrence formulas:

$$T_0 = 1,\ T_1 = x,\ T_{n+1} = 2xT_n - T_{n-1},\ n > 0. \tag{1}$$

Let us consider the principal properties of Chebyshev's polynomials.

1°. *The leading term of the polynomial T_{n+1} results from the leading term of the polynomial T_n ($n = 1, 2, \ldots$) multiplied by $2x$.*

□ We have

$$T_2 = 2x\cdot x - 1 = 2x^2 - 1,$$
$$T_3 = 2x(2x^2 - 1) - x = 4x^3 - 3x.$$

Therefore the leading term of the polynomial T_{n+1} is equal to $2^n x^{n+1}$. The general form of the polynomial T_{n+1} is

$$T_{n+1} = 2^n x^{n+1} + \dots . \blacksquare$$

2°. *All the polynomials* $T_{2n}(x)$ *are even functions and* $T_{2n+1}(x)$ *are odd functions.*

□ For $n = 0$ this is obvious. Let this be valid for a certain n. Then the function $2xT_{2n+1}(x)$ is even and, hence, $T_{2n+2} = 2xT_{2n+1}(x) - T_{2n}(x)$ is also an even function. Furthermore, the function $2xT_{2n+2}(x)$ is odd and therefore $T_{2n+3}(x) = 2xT_{2n+2}(x) - T_{2n+1}(x)$ is also an odd function. ■

3°. *If* $x \in [-1, 1]$, *then Chebyshev's polynomials have the following explicit expression:*

$$T_{n+1}(x) = \cos[(n+1)\arccos x], \quad n \geqslant -1. \tag{2}$$

□ We shall prove that the right-hand side of relation (2) satisfies definition (1) of Chebyshev's polynomials. Indeed,

$$\{\cos[(n+1)\arccos x]\}_{n=-1} = 1 = T_0,$$

$$\{\cos[(n+1)\arccos x]\}_{n=0} = x = T_1.$$

To prove that the recurrence formula holds true, we consider the obvious trigonometric relation

$$\cos(n+1)\theta = 2\cos\theta\cos n\theta - \cos(n-1)\theta.$$

Setting $\theta = \arccos x$, whence it follows that $x = \cos\theta$, we get $\cos[(n+1)\arccos x] = 2x\cos[n\arccos x] - \cos[(n-1)\arccos x]$. ■

4°. *On the interval* $[-1, 1]$ *the polynomials* $T_{n+1}(x)$ *have* $n+1$ *distinct roots:*

$$x_k = \cos\frac{2k+1}{2(n+1)}\pi \quad (k = 0, 1, \dots, n). \tag{3}$$

□ Using expression (2), we get the following equation for determining the roots of the polynomial $T_{n+1}(x)$:

$$(n+1)\arccos x_k = \frac{\pi}{2} + \pi k \quad (k = 0, 1, \dots, n).$$

Solving this equation for x_k, we arrive at relation (3). ■

5°. *On the interval* $[-1, 1]$ *the inequality*

$$|T_{n+1}(x)| \leqslant 1$$

holds true.

This follows immediately from relation (2).

From the same relation (2) we find all points x_m at which the polynomial $T_{n+1}(x)$ attains its extremal values ± 1. For this to occur, it is necessary that

$$(n+1)\arccos x_m = \pi m \quad (m = 0, 1, \dots, n+1)$$

and, consequently,

$$x_m = \cos\frac{m}{n+1}\pi \quad (m = 0, 1, \dots, n+1). \tag{5}$$

Substituting these values into relation (2), we obtain

$$T_{n+1}(x_m) = \cos m\pi = (-1)^m. \tag{6}$$

This means that the points, at which $T_{n+1} = 1$ and $T_{n+1} = -1$, alternate beginning with $x_0 = 1$, where $T_{n+1}(1) = 1$. Note once again that inequality (4) holds true not for all x. If $|x| > 1$, then arccos x does not exist on the set of real numbers.

Let us consider now polynomials

$$T_{n+1}(x) = 2^{-n}T_{n+1}(x) = x^{n+1} + \dots . \tag{7}$$

These are polynomials which have the smallest deviation from zero on the interval $[-1, 1]$. This fact is substantiated by the following theorem.

Theorem. *Let $P_{n+1}(x)$ be a polynomial of degree $n+1$ with the leading coefficient equal to 1. Then*

$$\max_{[-1,\,1]} |P_{n+1}(x)| \geqslant \max_{[-1,\,1]} |\overline{T}_{n+1}(x)| = 2^{-n}. \tag{8}$$

□ We assume that inequality (8) does not hold true. Then the polynomial $Q_n(x) = \overline{T}_{n+1} - P_{n+1}$ whose degree is not higher than n at all $n+2$ extremal points x_m of the polynomial T_{n+1} would coincide with the latter polynomial in sign and, consequently, would alternately assume positive and negative values at these points. Therefore $Q_n(x)$ must have $n+1$ distinct roots, and this is impossible for a polynomial of a degree not higher than n. The contradiction obtained proves the theorem. ■

Every interval $[a, b]$ can be obtained from the interval $[-1, 1]$ by means of a linear change of variables

$$x' = \frac{b+a}{2} + \frac{b-a}{2}\,x. \tag{9}$$

In this case the polynomial $\overline{T}_{n+1}(x)$ is transformed into a polynomial $\overline{T}_{n+1}\left(\frac{2x-(b+a)}{b-a}\right)$ with a leading coefficient $\left(\frac{2}{b-a}\right)^{n+1}$. Consequently,

$$\overline{T}^{[a,\,b]}_{n+1}(x) = (b-a)^{n+1}\,2^{-2n-1}T_{n+1}\left(\frac{2x-(b+a)}{b-a}\right) \tag{10}$$

is a polynomial with a leading coefficient 1 which has the smallest deviation from zero on the interval $[a, b]$, and the following inequality holds true for any polynomial $P_{n+1}(x)$ of degree $n+1$ with a leading coefficient 1:

$$\max_{[a,\,b]} |P_{n+1}(x)| \geqslant \max_{[a,\,b]} \left|\overline{T}^{[a,\,b]}_{n+1}(x)\right| = 2^{-n}\left(\frac{b-a}{2}\right)^{n+1}. \tag{11}$$

By virtue of the linear change of variables (9), the roots of the polynomial $\overline{T}_{n+1}^{[a,b]}(x)$ have the form

$$x_k=\frac{b+a}{2}+\frac{b-a}{2}\cos\frac{2k+1}{2(n+1)}\pi \quad (k=0, \ldots, n), \tag{12}$$

and the extremal points have the form

$$x_m=\frac{b+a}{2}+\frac{b-a}{2}\cos\frac{m}{n+1}\pi \quad (m=0, 1, \ldots, n+1). \tag{13}$$

We shall return now to the problem of the minimization of the error of interpolation Δ_1 on the interval $[a, b]$ for an arbitrary net on the class of $n+1$ times continuously differentiable functions which satisfy condition (7) from 7.5. We shall designate this class of functions as $C^{n+1}(M_{n+1}[a, b])$. By virtue of formula (8) from 7.5, to solve this problem we must minimize the quantity $\max\limits_{[a,b]}|\omega_n(x)|$. Since $\omega_n(x)$ is a polynomial of degree $n+1$ with a leading coefficient 1, it is evident that the quantity $\max\limits_{[a,b]}|\omega_n(x)|$ attains its minimum value for Chebyshev's polynomials $\overline{T}_{n+1}^{[a,b]}(x)$. Consequently, we must take as the interpolation nodes the points x_k defined by expression (12). In this case

$$\max_{[a,b]}|\omega_n(x)|=\max_{[a,b]}\left|\overline{T}_{n+1}^{[a,b]}(x)\right|=2^{-n}\left(\frac{b-a}{2}\right)^{n+1}, \tag{14}$$

and estimate (8) from 7.5 assumes the form

$$\Delta_1\leqslant\frac{M_{n+1}}{(n+1)!}2^{-n}\left(\frac{b-a}{2}\right)^{n+1}. \tag{15}$$

This estimate cannot be improved since we shall have an equality sign in it if we choose the polynomial of degree $n+1$

$$f(x)=\frac{M_{n+1}}{(n+1)!}x^{n+1}+a_nx^n+\ldots$$

as the function $f(x)$ and the points x_k defined by expression (12) as interpolation nodes.

Example 1. On the interval $[-1, 1]$ obtain a uniform estimate of the deviation of the function $f(x)=1-\cos(\pi x/2)$ from its interpolating polynomial constructed with the use of Chebyshev's nodes (3) for $n=2, 3, 4$.

△ Note first of all that for the function in question on the given interval we have $M_{n+1} = (\pi/2)^{n+1}$, $b - a = 2$. Therefore, by virtue of estimate (15), we have

$$n = 2, \quad \Delta_1 \leqslant \left(\frac{\pi}{2}\right)^3 \frac{1}{3!}\left(\frac{1}{2}\right)^2 \cong 0.17,$$

$$n = 3, \quad \Delta_1 \leqslant \left(\frac{\pi}{2}\right)^4 \frac{1}{4!}\left(\frac{1}{2}\right)^3 \cong 0.032,$$

$$n = 4, \quad \Delta_1 \leqslant \left(\frac{\pi}{2}\right)^5 \frac{1}{5!}\left(\frac{1}{2}\right)^4 \cong 0.005.$$

We recommend the reader to compare the solution obtained with the solution of Example 1 from 7.5. ▲

7.15. Interpolation with Multiple Nodes

Up till now we considered a problem in which the interpolation parameters, i.e. the coefficients of the interpolating polynomial, were defined only by the values of the interpolated function. This problem is often called an *interpolation problem in the sense of Lagrange* and the process of constructing the interpolating polynomial, *Lagrange's process*.

We shall consider now a more extensive problem, that of interpolation with respect to the values of a function and its derivatives or, as they also say, the *problem of multiple interpolation*.

Assume that the values f_i of a function f and its derivatives $f_i^{(k)}$ $(i = 0, 1, \ldots, m,\ k = 0, 1, \ldots, \alpha_i - 1)$ are given on the net Λ_m: $a \leqslant x_0 < x_1 < \ldots < x_m \leqslant b$ at the nodal points x_i, and, moreover, $\sum\limits_{i=0}^{m} \alpha_i = n + 1$. We have to construct a polynomial H_n whose values and derivatives up to the order $\alpha_i - 1$ at the nodes x_i $(i = 0, 1, \ldots, m)$ coincide with the values of f and its corresponding derivatives and to estimate the error.

Interpolation of this kind is the *interpolation in the sense of Hermite* and the corresponding polynomial H_n is the *Hermite polynomial.* The numbers α_i are the *multiplicities of the nodes* x_i. We can prove that the Hermite polynomial exists and is unique.

The remainder of the interpolation formula $f(x) \cong H_n(x)$ can be represented as

$$R_n(x) = f(x) - H_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{m} (x - x_i)^{\alpha_i},$$
$$\xi \in (a, b). \tag{1}$$

Assume now for definiteness that

$$| f^{(n+1)}(x) | \leqslant M_{n+1}, \quad x \in [a, b]. \tag{2}$$

Using this restriction and formula (1), we get the estimate of the error for a fixed point x:

$$\Delta_1 = | f(x) - H_n(x) | \leqslant \frac{M_{n+1}}{(n+1)!} \prod_{i=0}^{m} | x - x_i |^{\alpha_i}. \tag{3}$$

It is now easy to construct an estimate, uniform throughout the interval $[a, b]$, for the fixed net Λ_m. Indeed,

$$\Delta_1 = \max_{[a,b]} | R_n(x) | \leqslant \frac{M_{n+1}}{(n+1)!} \max_{[a,b]} | \Omega_n(x) |, \tag{4}$$

where

$$\Omega_n(x) = \prod_{i=0}^{m} (x - x_i)^{\alpha_i}. \tag{5}$$

Example. Construct the Hermite interpolating polynomial for the function $f = 1 - \cos(\pi x/2)$ using the nodes $x_0 = -1$, $x_1 = 0$, $x_2 = 1$ with multiplicities $\alpha_0 = 1$, $\alpha_1 = 2$ and $\alpha_2 = 1$ respectively. Get a uniform estimate of the error on the interval $[-1, 1]$.

△ We calculate the values of the function and of its derivative at the given nodes: $f(x_0) = f(x_2) = 1$, $f(x_1) = f'(x_1) = 0$. Then we construct the Hermite polynomial with due account of the multiplicities of the nodes:

$$H_3(x) = 1 \cdot \frac{x^2(x-1)}{-2} + 0 \cdot \frac{(x-1)(x+1)}{-1} + 1 \cdot \frac{(x+1)x^2}{2} = x^2.$$

Note that instead of a third-degree polynomial we have got a second-degree polynomial. This is a consequence of the symmetry of the initial data (but not of the function f).

Let us now find the estimate of the error. Using formula (4) and bearing in mind that $M_4 = (\pi/2)^4$ for the function being considered, we get

$$\Delta_1 \leqslant \left(\frac{\pi}{2}\right)^4 \cdot \frac{1}{4!} \cdot \max_{[-1,1]} |(x+1)x^2(x-1)|.$$

It is easy to show that $\max_{[-1,1]} | x + 1)x^2(x - 1) | = 0.25$. Therefore the final result is $\Delta_1 \leqslant 0.065$. ▲

7.16. Mathematical Apparatus of Trigonometric Interpolation

In the sections that follow we consider the problem of approximating functions by means of a trigonometric polynomial. This means that we take a linear combination of the trigonometric functions sin nx and cos nx as an approximating function.

To give a formal substantiation of the choice of trigonometric functions as approximating functions, we need some knowledge from the course of analysis concerning the Fourier series. Below we give some data on the Fourier series omitting the proofs of the statements.

Sequences. Series. Consider a function $f(x)$. We take the set of natural numbers as the domain of definition of this function, i.e. the argument x assumes the values 1, 2, . . ., n. A function of this kind is called a *sequence*.

A sequence is written in the form

$$a_1, a_2, \ldots, a_{n-1}, a_n, a_{n+1}, \ldots, \text{ or } \{a_n\}.$$

Here a_n is the *general term* of the sequence, a_{n-1} is a term preceding a_n and a_{n+1} is a term that succeeds a_n.

Here are some examples of sequences.

(1) The sequence 1, 2, 3, . . ., n, . . . whose general term $a_n = n$, is known as a *natural scale*.

(2) The sequence $a_1, a_2, \ldots, a_{n-1}, a_n, \ldots$, for which $a_n - a_{n-1} = d$, where d is a constant quantity, is an *arithmetic progression*. To define an arithmetic progression, it is sufficient to know its first term a_1 and the common difference d. Indeed, the general term is expressed by the formula

$$a_n = a_1 + d(n-1).$$

Since this formula can be used to find any term of the sequence by substituting the values $n = 1, 2, 3, \ldots$, the sequence is considered to be specified.

(3) The sequence $b_1, b_2, \ldots, b_{n-1}, b_n, \ldots$, for which $b_n = b_{n-1}q$, where q is a constant quantity, is a *geometric progression*. To define a geometric progression, it is sufficient to know its first term b_1 and the common ratio q. The general term is expressed by the formula

$$b_n = b_1 q^{n-1}.$$

(4) The sequence $c_1, c_2, \ldots, c_n, \ldots$, for which $c_n = c$, where c is a constant quantity, is a *constant sequence*.

(5) Here is another example of a sequence. We shall calculate the number e in succession with one, two, three etc. digits. We can represent the results of the calculation as

$$\begin{array}{ccccc} 2, & 2.7, & 2.71, & 2.718, & \ldots, \\ (1), & (2), & (3), & (4), & \ldots \end{array}$$

Labelling the values obtained by natural numbers, as it is shown in brackets, we get a sequence.

In addition to number series, we shall consider *functional sequences.*

Here are some examples of functional sequences:

$$(1)\ a_0,\ a_1x,\ a_2x^2,\ \ldots,\ a_{n-1}x^{n-1},\ \ldots,$$

$$(2)\ \sin x,\ \sin 2x,\ \sin 3x,\ \ldots,\ \sin nx,\ \ldots$$

Recall the definition of the limit of a number sequence. The number A is the *limit of the sequence* $\{a_n\}$ if for any $\varepsilon > 0$ there is a number N such that the inequality $|a_n - A| < \varepsilon$ is satisfied for all $n > N$. Then we write $\lim\limits_{n\to\infty} a_n = A$.

A sequence which has a limit is a *convergent sequence*, otherwise it is a *divergent sequence.*

Example 1. Show that the geometric progression $b, bq, bq^2, \ldots, bq^{n-1}, \ldots$ is a convergent sequence for $|q| < 1$ and a divergent sequence for $|q| \geqslant 1$.

△ (1) We shall first consider the case when $|q| < 1$. We shall show that

$$\lim_{n\to\infty} bq^{n-1} = 0,$$

i.e. proceeding from the given $\varepsilon > 0$, we shall find N such that the inequality $|bq^{n-1} - 0| < \varepsilon$ is satisfied for $n > N$. To do this, we solve the inequality for n. We rewrite it in the form

$$|b|\,|q^{n-1} < \varepsilon, \quad \text{or} \quad |q|^{n-1} < \varepsilon/|b|.$$

We take logarithms of the last inequality:

$$(n-1)\ln|q| < \ln(\varepsilon/|b|).$$

Dividing both its sides by the negative number $\ln|q|$, we get

$$n-1 > \frac{\ln(\varepsilon/|b|)}{\ln|q|}, \quad \text{i.e.} \quad n > 1 + \frac{\ln(\varepsilon/|b|)}{\ln|q|}.$$

Evidently, it is sufficient to take

$$N = E\left(1 + \frac{\ln(\varepsilon/|b|)}{\ln|q|}\right),$$

where $E(x)$ is the greatest integer not exceeding x, as N.

(2) Let us now consider the case when $q > 1$, $b > 0$. We shall show that the sequence is divergent. To do this, it is sufficient to show that for any arbitrarily large M there is N such that the inequality $bq^{n-1} > M$ is satisfied for all $n > N$. Solving this inequality for n, we get

$$n > 1 + \frac{\ln(M/b)}{\ln q}.$$

We take

$$N = E\left(1 + \frac{\ln(M/b)}{\ln q}\right)$$

as N.

Note that for $q=1$ the sequence is constant and $\lim\limits_{n\to\infty} bq^{n-1}=b$. We can show that in all other cases the sequence is divergent. ▲

Consider a sequence $a_1, a_2, \ldots, a_n, \ldots$. An expression of the form

$$a_1+a_2+\ldots+a_n+\ldots=\sum_{n=1}^{\infty} a_n$$

is a *series*; here a_n is the nth term of the series.

The sum of the first n terms is its nth *partial sum*:

$$S_n=\sum_{i=1}^{n} a_i.$$

The *sum of a series* is the limit of the sequence of its partial sums:

$$\lim_{n\to\infty} S_n=S.$$

If a series has a sum, it is *convergent*, otherwise it is *divergent*.

Here are some examples of convergent and divergent series.

Example 2. We consider a series resulting from an arithmetic progression and write its nth partial sum

$$S_n=\frac{a_1+a_n}{2}\cdot n=\frac{2a_1+d(n-1)}{2}\cdot n=\frac{d}{2}n^2+\frac{2a_1-d}{2}n.$$

Evidently, as n tends to infinity, its partial sum increases indefinitely in absolute value and, consequently, the series diverges.

Example 3. We consider a series resulting from a geometric progression for $|q|<1$ and find its sum. We use the formula for the sum of n terms of the geometric progression:

$$S_n=\frac{b_1(1-q^n)}{1-q}=\frac{b_1}{1-q}-\frac{b_1}{1-q}q^n.$$

We have proved (see Example 1) that $\lim\limits_{n\to\infty} q^n=0$ $(|q|<1)$, consequently,

$$S=\lim_{n\to\infty} S_n=\frac{b_1}{1-q}.$$

Series whose terms are functions are *functional series*. Here are examples of functional series:

$$a_0+a_1x+a_2x^2+\ldots+a_{n-1}x^{n-1}+\ldots,$$

$$b_0+b_1\cos x+b_2\cos 2x+\ldots+b_{n-1}\cos(n-1)x+\ldots.$$

The first of them is a *power series* and the second is a *trigonometric series*.

Let us consider a functional series

$$u_1(x)+u_2(x)+\ldots+u_n(x)+\ldots,$$

where $u_n(x)$ are functions defined on the interval $[a, b]$. Let $x_0 \in [a, b]$, and then the series

$$u_1(x_0) + u_2(x_0) + \ldots + u_n(x_0) + \ldots$$

is a number series and may prove to be convergent as well as divergent.

The collection of all values of $x \in [a, b]$, for which the corresponding number series converges, is the *domain of convergence* of the functional series.

It is evident that

$$S(x) = \lim_{n \to \infty} S_n(x),$$

where $S_n(x) = \sum_{i=1}^{n} u_i(x)$ depends on the choice of the variable x, i.e. the sum $S(x)$ of the functional series is a function of the point x.

Let $\{S_n(x)\}$ be a sequence of partial sums of a functional series defined on the same closed interval $[a, b]$. A functional series is *uniformly convergent* to the function $S(x)$, defined on $[a, b]$, if any $\varepsilon > 0$ can be associated with a number N, independent of $x \in [a, b]$, such that the inequality $|S_n(x) - S(x)| < \varepsilon$ is satisfied for any $n > N$.

Consider an example which illustrates the difference between the concepts of convergence and uniform convergence of a series.

Example 4. For the series $\sum_{n=1}^{\infty} \frac{x}{(1+x)^n}$, where $0 \leqslant x \leqslant 1$, the partial sum is

$$S_n(x) = \sum_{i=1}^{n} \frac{x}{(1+x)^i} = 1 - \frac{1}{(1+x)^n}.$$

We shall show that the series converges on the interval being considered.

Indeed, if $x = 0$, then $S_n(0) = S(0) = 0$. Assume now that $x > 0$. We shall prove that in this case $S(x) = 1$, i.e. using the given $\varepsilon > 0$, we shall find N such that the inequality

$$\left| 1 - \frac{1}{(1+x)^n} - 1 \right| < \varepsilon$$

is satisfied for $n > N$. Solving this inequality, we get

$$N = E\left(-\frac{\ln \varepsilon}{\ln(1+x)} \right).$$

We can see from the last relation that in general N depends not only on ε but also on x and increases indefinitely for one and the same ε for x tending to zero. This means that proceeding from the given ε, we cannot choose a unique N for all $x \in [0, 1]$, in other words, the series is *not uniformly convergent on the indicated interval.*

This example shows that a sequence of partial sums continuous on a closed interval may converge to a function discontinuous on that interval. One of the reasons for the introduction of the concept of a uniform convergence of a series is that a uniformly convergent series of continuous functions has a continuous function as its sum.

Expansion of functions in Fourier series. Many problems in science and engineering are connected with periodic functions which reflect cyclic processes.

The function $f(x)$ is *periodic* with period $T > 0$ if it satisfies the equality

$$f(x) = f(x + T). \tag{1}$$

For practical purposes it is convenient to represent functions of this kind as trigonometric series or their partial sums with sufficient accuracy.

A functional series of the form

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx) \tag{2}$$

is *trigonometric*, where a_n and b_n are real numbers independent of x.

Assume that this series converges for any x from the interval $[-\pi, \pi]$, and then it defines a periodic function $f(x)$ with period $T = 2\pi$.

A series of form (2) is known as a *Fourier series* for the function $f(x)$ integrable on the interval $[-\pi, \pi]$ if its coefficients can be calculated from the following formulas:

$$a_0 = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x)\, dx, \tag{3}$$

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx\, dx \quad (n = 1, 2, \ldots), \tag{4}$$

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx\, dx \quad (n = 1, 2, \ldots). \tag{5}$$

We can formally consider the Fourier series for the function $f(x)$. However, the following questions arise: (1) whether the Fourier series converges for the function $f(x)$, (2) if the series converges, then whether it has $f(x)$ as its sum. Dirichlet's theorem gives answers to these questions. Before formulating this theorem, we shall recall some concepts.

The function $f(x)$ is *monotonic* on an interval if for any x_1 and x_2, which belong to this interval and are such that $x_1 < x_2$, only one of the inequalities $f(x_1) \leqslant f(x_2)$ and $f(x_1) \geqslant f(x_2)$ is satisfied.

The function $f(x)$ is *piecewise-monotonic* on an interval if the interval can be divided into a finite number of open intervals in each of which the function is monotonic.

The function $f(x)$ is *piecewise-continuous* on an interval if it has a finite number of points of discontinuity on that interval.

We designate the limit of the function $f(x)$ when x tends to a from the right (right-hand limit) as $f(a+0)$ and, respectively, the left-hand limit as $f(a-0)$.

Dirichlet's theorem. *If the function $f(x)$ given on the interval $[-\pi, \pi]$ is piecewise-monotonic and piecewise-continuous, then the Fourie. series of this function converges throughout the interval $[-\pi, \pi]$ and its sum is equal to*

(1) *$f(x)$ at all points of continuity belonging to $(-\pi, \pi)$,*

(2) *$\frac{1}{2}[f(x-0)+f(x+0)]$ at all points of discontinuity belonging to $(-\pi, \pi)$,*

(3) *$\frac{1}{2}[f(-\pi+0)+f(\pi-0)]$ at the endpoints of the interval, i.e. at the points $x = -\pi$* and $x=\pi$.

In what follows we shall write that

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx) \qquad (6)$$

in the sense of Dirichlet's theorem.

Dirichlet's theorem does not assert the uniform convergence of the Fourier series to the function $f(x)$. However, if we strengthen the properties which the function must satisfy, i.e. require that it should be continuous throughout the interval $[-\pi, \pi]$, piecewise-monotonic on it and that the equality $f(-\pi) = f(\pi)$ should be satisfied, then the Fourier series for such a function will converge uniformly to the function $f(x)$ throughout the interval $[-\pi, \pi]$.

We can show that *for an even function all the coefficients b_n are zero and the corresponding Fourier series does not include sines*:

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos nx, \qquad (7)$$

where

$$a_n = \frac{2}{\pi} \int_0^{\pi} f(x) \cos nx \, dx \quad (n=0, 1, 2, \ldots). \qquad (8)$$

Similarly, *for an odd function all the coefficients a_n are zero and the corresponding Fourier series does not include cosines*'

$$f(x) = \sum_{n=1}^{\infty} b_n \sin nx, \qquad (9)$$

where

$$b_n=\frac{2}{\pi}\int\limits_0^{\pi} f(x)\sin nx\,dx \quad (n=1,\ 2,\ \ldots). \tag{10}$$

Example 1. Represent the function

$$f(x)=\begin{cases} -x & \text{for } -\pi\leqslant x\leqslant 0,\\ 2x & \text{for } \quad 0<x\leqslant\pi\end{cases}$$

as a Fourier series.

△ We find the Fourier coefficients of the function $f(x)$. From formulas (3) and (4) we find the coefficients a_0 and a_n:

$$a_0=\frac{1}{\pi}\int\limits_{-\pi}^{0}(-x)\,dx+\frac{1}{\pi}\int\limits_0^{\pi}2x\,dx=\frac{1}{\pi}\cdot\frac{\pi^2}{2}+\frac{1}{\pi}\cdot\pi^2=\frac{3}{2}\pi,$$

$$a_n=\frac{1}{\pi}\int\limits_{-\pi}^{0}(-x)\cos nx\,dx+\frac{1}{\pi}\int\limits_0^{\pi}2x\cos nx\,dx.$$

Integrating by parts, we obtain

$$a_n=-\frac{1}{\pi}\left(\frac{1}{n}x\sin nx\,\Big|_{-\pi}^{0}-\frac{1}{n}\int\limits_{-\pi}^{0}\sin nx\,dx\right.$$

$$\left.-\frac{2}{n}x\sin nx\,\Big|_0^{\pi}+\frac{2}{n}\int\limits_0^{\pi}\sin nx\,dx\right)$$

$$=-\frac{1}{\pi n}\left(\frac{1}{n}\cos nx\,\Big|_{-\pi}^{0}-\frac{2}{n}\cos nx\,\Big|_0^{\pi}\right)=-\frac{3}{\pi n^2}[1-(-1)^n],$$

i.e.

$$a_{2k}=0,\quad a_{2k-1}=-\frac{6}{\pi(2k-1)^2}\quad (k=1,\ 2,\ 3,\ \ldots).$$

The coefficients b_n can be found from formula (5):

$$b_n=\frac{1}{\pi}\int\limits_{-\pi}^{0}(-x)\sin nx\,dx+\frac{1}{\pi}\int\limits_0^{\pi}2x\sin nx\,dx$$

$$=-\frac{1}{\pi}\left(-\frac{x}{n}\cos nx\,\Big|^{0}+\frac{1}{n}\int^{\pi}\cos nx\,dx\right.$$

$$+\frac{2x}{n}\cos nx\Big|_0^{\pi}-\frac{2}{n}\int_0^{\pi}\cos nx\,dx\Big)$$

$$=-\frac{1}{\pi}\left[-\frac{\pi}{n}(-1)^n+2\frac{\pi}{n}(-1)^n\right]=\frac{(-1)^{n+1}}{n},$$

i.e.

$$b_{2k}=-\frac{1}{2k},\quad b_{2k-1}=\frac{1}{2k-1}\quad (k=1,\ 2,\ 3,\ \ldots).$$

Thus the Fourier series of this function has the form

$$f(x)=\frac{3}{4}\pi-\frac{6}{\pi}\sum_{k=1}^{\infty}\frac{1}{(2k-1)^2}\cos(2k-1)x$$

$$+\sum_{k=1}^{\infty}\frac{1}{2k-1}\sin(2k-1)x-\frac{1}{2}\sum_{k=1}^{\infty}\frac{1}{k}\sin 2kx.$$

In the interval $(-\pi, \pi)$ the series converges to the function $f(x)$ and at the points $x=\pm\pi$ to the number

$$\frac{1}{2}\cdot[f(-\pi+0)+f(\pi-0)]=\frac{3}{2}\pi. \ \blacktriangle$$

Example 2. Represent the function

$$f(x)=\begin{cases}-\sin x & \text{for } -\pi\leqslant x\leqslant 0,\\ \sin x & \text{for } 0<x\leqslant\pi\end{cases}$$

as a Fourier series.

$\triangle$ This function is even and, consequently, all the coefficients $b_n=0$ and a_n can be found from formula (8):

$$a_n=\frac{2}{\pi}\int_0^{\pi}\sin x\cos nx\,dx\quad (n=0,\ 1,\ 2,\ \ldots).$$

From this we have

$$a_0=\frac{2}{\pi}\int_0^{\pi}\sin x\,dx=\frac{4}{\pi}.$$

Furthermore,

$$a_n=\frac{2}{\pi}\int_0^{\pi}\sin x\cos nx\,dx=\frac{1}{\pi}\int_0^{\pi}[\sin(n+1)x-\sin(n-1)x]\,dx$$

$$=\begin{cases}0 & \text{for } n=2k-1,\\ -\dfrac{4}{\pi(n^2-1)} & \text{for } n=2k\ (k=1,\ 2,\ 3,\ \ldots).\end{cases}$$

Consequently,

$$f(x) = |\sin x| = \frac{2}{\pi} - \frac{4}{\pi} \sum_{k=1}^{\infty} \frac{1}{4k^2-1} \cos 2kx.$$

Note that the series obtained converges to the function $|\sin x|$ throughout the interval $[-\pi, \pi]$. ▲

Example 3. Represent the function

$$f(x) = \begin{cases} -1 & \text{for } -\pi < x < 0, \\ 0 & \text{for } \quad x = 0, \\ 1 & \text{for } \quad 0 < x < \pi \end{cases}$$

as a Fourier series.

△ This function is odd and, consequently, all the coefficients $a_n = 0$ and b_n can be found from formula (10):

$$b_n = \frac{2}{\pi} \int_0^{\pi} 1 \cdot \sin nx \, dx = -\frac{2}{\pi n} \cos nx \Big|_0^{\pi} = \frac{2}{\pi n} [1 - (-1)^n].$$

Thus all the even coefficients b_n are zero and the odd ones have the form $b_{2k-1} = \frac{4}{\pi (2k-1)}$. Consequently,

$$f(x) = \frac{4}{\pi} \sum_{k=1}^{\infty} \frac{1}{2k-1} \sin (2k-1) x.$$

Evidently, at the points $x = 0$ and $x = \pm \pi$ the sum of the series is equal to zero. ▲

Example 4. Represent the function

$$f(x) = \begin{cases} x & \text{for } 0 < x < 1, \\ 1 & \text{for } 1 < x < 2 \end{cases}$$

as a Fourier series.

△ Since the function is defined in an interval different from $(-\pi, \pi)$, we make a change of the independent variable using the formula $x = (x' + \pi)/\pi$, or $x' = \pi (x - 1)$. Thus we get a function

$$f(x') = \begin{cases} \frac{1}{\pi}(x' + \pi) & \text{for } -\pi < x' \leqslant 0, \\ 1 & \text{for } \quad 0 < x' < \pi. \end{cases}$$

Since this function is defined in the interval $(-\pi, \pi)$, we can write a Fourier series for it. We calculate the coefficients of this series:

$$a_0 = \frac{1}{\pi} \int_{-\pi}^{0} \frac{x' + \pi}{\pi} \, dx' + \frac{1}{\pi} \int_0^{\pi} 1 \cdot dx' = \frac{3}{2},$$

$$a_n = \frac{1}{\pi} \int^{0} \frac{x' + \pi}{\pi} \cos nx' \, dx' + \frac{1}{\pi} \int^{\pi} 1 \cdot \cos nx' \, dx'$$

$$=\frac{1}{n\pi^2}\int_{-\pi}^{0}\sin nx'\,dx'=\frac{1}{\pi^2n^2}[1-(-1)^n],$$

$$b_n=\frac{1}{\pi}\int_{-\pi}^{0}\frac{x'+\pi}{\pi}\sin nx'\,dx'+\frac{1}{\pi}\int_{0}^{\pi}1\cdot\sin nx'\,dx'$$

$$=\frac{1}{\pi^2}\int_{-\pi}^{0}x'\sin nx'\,dx'=-\frac{(-1)^n}{\pi n}.$$

Consequently,

$$f(x)=\frac{3}{4}+\frac{2}{\pi^2}\sum_{k=1}^{\infty}\frac{1}{(2k-1)^2}\cos[(2k-1)\pi(x-1)]$$

$$-\frac{1}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^n}{n}\sin[n\pi(x-1)].$$

We calculate the values of the sum of the series at the endpoints of the interval:

$$\frac{1}{2}[f(0+0)+f(2-0)]=\frac{0+1}{2}=\frac{1}{2}.$$

The result obtained makes it possible to find the sum of the number series

$$S=\sum_{k=1}^{\infty}\frac{1}{(2k-1)^2}=1+\frac{1}{3^2}+\frac{1}{5^2}+\dots+\frac{1}{(2k-1)^2}+\dots.$$

Indeed, on the basis of Dirichlet's theorem, for $x=0$ or $x=2$, we have

$$\frac{1}{2}=\frac{3}{4}-\frac{2}{\pi^2}\sum_{k=1}^{\infty}\frac{1}{(2k-1)^2},$$

whence it follows that

$$\sum_{k=1}^{\infty}\frac{1}{(2k-1)^2}=\frac{\pi^2}{8}. \blacktriangle$$

It should be pointed out in conclusion that the integral of the function $f(x)$ results from a term-by-term integration of the corresponding Fourier series and the derivative $f'(x)$ can be obtained by means of a term-by-term differentiation. The condition $f(-\pi)=f(\pi)$ is obligatory in differentiation.

7.17. Trigonometric Interpolation

The operation of representing the function $f(x)$ as a Fourier series is known as a *harmonic*, or *Fourier*, *analysis*. In practice, we have to restrict our computations to the first few terms of a Fourier series. As a result we get only an approximate analytic expression for the function $f(x)$ in the form of a trigonometric polynomial of order N:

$$Q_N(x) = \frac{a_0}{2} + \sum_{n=1}^{N} (a_n \cos nx + b_n \sin nx)$$
$$(-\pi \leqslant x \leqslant \pi). \tag{1}$$

Beside this, formulas (3)-(5) from 7.16 are suitable for calculation of Fourier coefficients only in the case of an analytic representation of the function. In practice, as a rule, the function $f(x)$ is given in tabular form or as a graph. Therefore a problem arises of an approximation of the Fourier coefficients with the use of a finite number of the available values of the function.

Generalizing all we have said above, we shall formulate a problem of the numerical, or, as it is also called, harmonic analysis: we have to approximate, *in the interval* $(0, T)$ by the trigonometric polynomial of order N, the function $y = f(x)$, for which we know m of its values $y_k = f(x_k)$ for $x_k = kT/m$ $(k = 0, 1, 2, \ldots, m-1)$.

A trigonometric polynomial for the function defined on the interval $(0, T)$ has the form

$$Q_N(x) = \frac{a_0}{2} + \sum_{n=1}^{N} \left(a_n \cos n \frac{2\pi}{T} x + b_n \sin n \frac{2\pi}{T} x\right). \tag{2}$$

The coefficients a_n and b_n are defined by the following relations:

$$a_n = \frac{2}{T} \int_0^T f(x) \cos n \frac{2\pi}{T} x \, dx, \tag{3}$$

$$(n = 0, 1, 2, \ldots, N).$$

$$b_n = \frac{2}{T} \int_0^T f(x) \sin n \frac{2\pi}{T} x \, dx, \tag{4}$$

Employing, in (3) and (4), the rectangular formula for calculating the integrals from the values of the integrands at the points $x_k = kT/m$ ($k = 0, 1, 2, \ldots, m-1$), we have

$$a_n = \frac{2}{m} \sum_{k=0}^{m-1} y_k \cos n \frac{2\pi k}{m}, \tag{5}$$

$$(n = 0, 1, 2, \ldots, N).$$

$$b_n = \frac{2}{m} \sum_{k=0}^{m-1} y_k \sin n \frac{2\pi k}{m}. \tag{6}$$

Thus the trigonometric polynomial (2) whose coefficients a_n and b_n can be found from formulas (5) and (6), serves as a solution of the problem posed.

We can show that for $m > 2N$ polynomial (2) is the best approximation of the function $f(x)$ in the sense of the method of the least squares if its coefficients are calculated from formulas (5) and (6). To put it otherwise, coefficients (5) and (6) minimize the sum of the squares of the deviations

$$\delta_N^2 = \sum_{k=0}^{m-1} [Q_N(x_k) - y_k]^2. \tag{7}$$

In a special case, when $m = 2N$, the coefficients a_n and b_n ($n = 0, 1, 2, \ldots, N-1$) are defined by relations (5) and (6) and the coefficient a_N is

$$a_N = \frac{1}{m} \sum_{k=0}^{m-1} (-1)^k y_k. \tag{8}$$

The polynomial $Q_N(x)$ itself becomes an interpolating polynomial since in this case, for any b_N, there hold relations $Q_N(x_k) = y_k$ for all $x_k = kT/m$ ($k = 0, 1, 2, \ldots, m-1$).

Example. We investigate the dynamics of the production of sugar from sugar-beet. This production is of a periodic nature which is due to the periodicity of growth and the conditions of preservation of raw material. Therefore, we can take the trigonometric polynomial (2) for $m = 12$ as a function approximating the dynamics of sugar production. (This corresponds to the number of months in an annual cycle and makes it possible to reveal the peculiarity of the production, its seasonal nature.) Consequently,

$$Q_N(x)=\frac{a_0}{2}+\sum_{n=1}^{N}\left(a_n\cos n\frac{\pi}{6}x+b_n\sin n\frac{\pi}{6}x\right)\quad(0\leqslant x\leqslant 11).$$

In economic investigations, not more than four harmonics are usually chosen for a good approximation of a dynamic periodic series.

The expressions for the coefficients a_n and b_n have the form

$$a_n=\frac{1}{6}\sum_{h=0}^{11}y_h\cos n\frac{\pi}{6}x_h,\quad b_n=\frac{1}{6}\sum_{h=0}^{11}y_h\sin n\frac{\pi}{6}x_h.$$

We calculate these coefficients for the first four harmonics of the polynomial $Q_N(x)$ and tabulate the necessary computations (see Table 7.13).

From this table we find that $a_0=108$, $a_1=34.99$, $a_2=7.75$, $a_3=-3$, $a_4=-1.25$, $b_1=-6.11$, $b_2=-11.98$, $b_3=-4$, $b_4=1.59$. Thus we have the following four mathematical models of the seasonal nature of sugar production:

$$Q_1(x)=54+34.99\cos\frac{\pi}{6}x-6.11\sin\frac{\pi}{6}x,$$

$$Q_2(x)=54+34.99\cos\frac{\pi}{6}x-6.11\sin\frac{\pi}{6}x+7.75\cos\frac{\pi}{3}x$$
$$-11.98\sin\frac{\pi}{3}x,$$

$$Q_3(x)=54+34.99\cos\frac{\pi}{6}x-6.11\sin\frac{\pi}{6}x+7.75\cos\frac{\pi}{3}x$$
$$-11.98\sin\frac{\pi}{3}x-3\cos\frac{\pi}{2}x-4\sin\frac{\pi}{2}x,$$

$$Q_4(x)=54+34.99\cos\frac{\pi}{6}x-6.11\sin\frac{\pi}{6}x+7.75\cos\frac{\pi}{3}x$$
$$11.98\sin\frac{\pi}{3}x-3\cos\frac{\pi}{2}x-4\sin\frac{\pi}{2}x$$
$$-1.25\cos\frac{2\pi}{3}x+1.59\sin\frac{2\pi}{3}x.$$

When we compare $Q_i(x_h)$ with the respective values of y_h, we see that already the first harmonic yields, in general, a correct model of the dynamics of sugar production, indicating its seasonal nature.

Table 7.13

Months	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	$\frac{1}{6}\Sigma$
x_k	0	1	2	3	4	5	6	7	8	9	10	11	
Output in convent. units	95	71	55	43	36	31	28	26	25	45	91	102	108
$\cos\frac{\pi}{6}x_k$	1	0.866	0.5	0	−0.5	−0.866	−1	−0.866	−0.5	0	0.5	0.866	
$\sin\frac{\pi}{6}x_k$	0	0.5	0.866	1	0.866	0.5	0	−0.5	−0.866	−1	−0.866	−0.5	
$\cos\frac{\pi}{3}x_k$	1	0.5	−0.5	−1	−0.5	0.5	1	0.5	−0.5	−1	−0.5	0.5	
$\sin\frac{\pi}{3}x_k$	0	0.866	0.866	0	−0.866	−0.866	0	0.866	0.866	0	−0.866	−0.866	
$\cos\frac{\pi}{2}x_k$	1	0	−1	0	1	0	−1	0	1	0	−1	0	
$\sin\frac{\pi}{2}x_k$	0	1	0	−1	0	1	0	−1	0	1	0	−1	
$\cos\frac{2\pi}{3}x_k$	1	−0.5	−0.5	1	−0.5	−0.5	1	−0.5	−0.5	1	−0.5	−0.5	

$\sin \frac{2\pi}{3} x_k$	0	0.866	−0.866	0	0.866	−0.866	0	0.866	−0.866	0	0.866	−0.866	
$y_k \cos \frac{\pi}{6} y_k$	95	61.49	27.5	0	−18	−26.85	−28	−25.52	−12.5	0	45.5	83.33	34.99
$y_k \sin \frac{\pi}{6} y_k$	0	35.5	47.63	43	31.18	15.5	0	−13	21.65	−45	−78.81	−51	−6.11
$y_k \cos \frac{\pi}{3} y_k$	95	35.5	−27.5	−43	−18	15.5	28	13	−12.5	−45	−45.5	51	7.75
$y_k \sin \frac{\pi}{3} y_k$	0	61.49	47.63	0	31.18	−28.85	0	22.52	21.65	0	−78.81	−88.33	−11.98
$y_k \cos \frac{\pi}{2} y_k$	95	0	−55	0	36	0	−28	0	25	0	−91	0	−3
$y_k \sin \frac{\pi}{2} y_k$	0	71	0	−43	0	31	0	−26	0	45	0	−102	−4
$y_k \cos \frac{2\pi}{3} y_k$	95	−35.5	−27.5	43	−18	−15.5	28	−13	−12.5	45	−45.5	−51	−1.25
$y_k \sin \frac{2\pi}{3} y_k$	0	61.49	−47.63	0	31.18	−26.85	0	22.52	−21.65	0	78.81	−88.33	1.59

Let us calculate the mean quadratic deviations $\delta_l = \sqrt{\sum_{k=0}^{I} [Q_l(x_k) - y_k]^2}$ for all $Q_l(x)$. We find that $\delta_1 = 37.80$, $\delta_2 = 14.40$, $\delta_3 = 7.59$, $\delta_4 = 5.75$. As would be expected, the values of δ_l decrease monotonically with an increase in l, δ_4 differing but little from δ_3. In addition, the values of δ_3 and δ_4 themselves are small and therefore the polynomial $Q_3(x)$ is already a close approximation of the series which characterizes the annual dynamics of sugar production.

7.18. Numerical Methods of Determining the Fourier Coefficients

Consider a Fourier series converging to the periodic function $f(x)$:

$$f(x) = \frac{a_0}{2} + \sum_{m=1} (a_m \cos mx + b_m \sin mx), \qquad (1)$$

where

$$a_m = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos mx \, dx \quad (m = 0, 1, 2, \ldots), \qquad (2)$$

$$b_m = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin mx \, dx \quad (m = 1, 2, 3, \ldots). \qquad (3)$$

In the preceding section we formulated a problem of approximation of the function $f(x)$ by the trigonometric polynomial $Q_N(x)$. We used there the rectangular formula to calculate the coefficients a_m and b_m with the aid of integrals.

In the general case, the approximation of the coefficients a_m and b_m is based on the replacement of the integrals in formulas (3) and (4) from 7.17 by their values obtained from one of the formulas of approximate integration. In this section we shall use the trapezoid rule.

We assume that the function $f(x)$ is periodic with period 2π. Note that when determining the coefficients a_m and b_m, we can consider any integration interval 2π long rather than the ordinary integration limits from $-\pi$

to π. To make the calculations more convenient, we take an interval from 0 to 2π so that

$$a_m = \frac{1}{\pi}\int_0^{2\pi} f(x)\cos mx\,\mathrm{d}x \quad (m = 0, 1, 2, \ldots), \tag{4}$$

$$b_m = \frac{1}{\pi}\int_0^{2\pi} f(x)\sin mx\,\mathrm{d}x \quad (m = 1, 2, 3, \ldots). \tag{5}$$

Dividing the interval $[0, 2\pi]$ into N equal parts, we get division points $0,\ 1\cdot\frac{2\pi}{N},\ 2\cdot\frac{2\pi}{N},\ \ldots,\ (N-1)\frac{2\pi}{N}$, 2π. We designate the corresponding values of the function $f(x)$ at the division points as $y_0, y_1, y_2, \ldots, y_{N-1}, y_N = y_0$. Applying the trapezoid rule, we get the following approximate formulas for the calculation of the coefficients a_m and b_m:

$$\frac{N}{2}a_0 = \sum_{h=0}^{N-1} y_h = y_0 + y_1 + \ldots + y_{N-1},$$

$$\frac{N}{2}a_m = \sum_{h=0}^{N-1} y_h \cos k\frac{2m\pi}{N} = y_0 + y_1\cos\frac{2m\pi}{N} + \ldots + y_{N-1}\cos(N-1)\frac{2m\pi}{N},$$

$$\frac{N}{2}b_m = \sum_{h=1}^{N-1} y_h \sin k\frac{2m\pi}{N} = y_1\sin\frac{2m\pi}{N} + \ldots + y_{N-1}\sin(N-1)\frac{2m\pi}{N}.$$

Assume that $N = 12$, i.e. the interval $[0, 2\pi]$ is divided into 12 equal parts so that we use the values of the argument $0, \pi/6, \pi/3, \ldots, 11\pi/6$, associated with the values of the function $y_0, y_1, y_2, \ldots, y_{11}$, and the values by which these values are multiplied are ± 1, $\pm\sin(\pi/6) = \pm 0.5$, $\pm\sin(\pi/3) = \pm 0.866$, From this, omitting cumbersome computations, we obtain

$$6a_0 = y_0 + y_1 + y_2 + y_3 + y_4 + y_5 + y_6 + y_7 + y_8 + y_9 + y_{10} + y_{11},$$

$$6a_1 = (y_2 + y_{10} - y_4 - y_8) \sin \frac{\pi}{6} + (y_1 + y_{11} - y_5 - y_7) \sin \frac{\pi}{3} + (y_0 - y_6),$$

$$6a_2 = (y_1 + y_5 + y_7 + y_{11} - y_2 - y_4 - y_8 - y_{10}) \sin \frac{\pi}{6} + (y_0 + y_6 - y_3 - y_9),$$

$$6a_3 = y_0 + y_4 + y_8 - y_2 - y_6 - y_{10},$$

$$6b_1 = (y_1 + y_5 - y_7 - y_{11}) \sin \frac{\pi}{6} + (y_2 + y_4 - y_8 - y_{10}) \sin \frac{\pi}{3} + (y_3 - y_9),$$

$$6b_2 = (y_1 + y_2 + y_7 + y_8 - y_4 - y_5 - y_{10} - y_{11}) \sin \frac{\pi}{3},$$

$$6b_3 = y_1 + y_5 + y_9 - y_3 - y_7 - y_{11},$$

and so on.

To minimize the number of arithmetic operations necessary to obtain the values of a_m and b_m, we use a special computing scheme known as **Runge's scheme**.

1st step. We write the values of the function $f(x)$ in the following order:

$$\begin{array}{ccccccc} y_0 & y_1 & y_2 & y_3 & y_4 & y_5 & y_6 \\ & y_{11} & y_{10} & y_9 & y_8 & y_7 & \end{array}$$

2nd step. We calculate the sums and the differences of each pair of values which are under each other and write the resulting sums and differences as follows:

$$\begin{array}{rccccccc} & y_0 & y_1 & y_2 & y_3 & y_4 & y_5 & y_6 \\ & & y_{11} & y_{10} & y_9 & y_8 & y_7 & \\ \hline \text{sums} & u_0 & u_1 & u_2 & u_3 & u_4 & u_5 & u_6 \\ \text{differences} & & v_1 & v_2 & v_3 & v_4 & v_5 & \end{array} \qquad (6)$$

3rd step. We perform similar operations with the sums and differences (6):

$$\begin{array}{rcccc} & u_0 & u_1 & u_2 & u_3 \\ & u_6 & u_5 & u_4 & \\ \hline \text{sums} & c_0 & c_1 & c_2 & c_3 \\ \text{differences} & d_0 & d_1 & d_2 & \end{array} \qquad \begin{array}{ccc} v_1 & v_2 & v_3 \\ v_5 & v_4 & \\ \hline g_1 & g_2 & g_3 \\ h_1 & h_2 & \end{array} \qquad (7)$$

4th step. We calculate the values of a_m and b_m using the approximate formulas

$$\left.\begin{aligned} 6a_0 &= c_0 + c_1 + c_2 + c_3,\\ 6a_1 &= d_0 + 0.866d_1 + 0.5d_2,\\ 6a_2 &= (c_0 - c_3) + 0.5(c_1 - c_2),\\ 6a_3 &= d_0 - d_2,\\ 6b_1 &= 0.5g_1 + 0.866g_2 + g_3,\\ 6b_2 &= 0.866(h_1 - h_2),\\ 6b_3 &= g_1 - g_3, \end{aligned}\right\} \qquad (8)$$

and so on.

To compare the coefficients a_m and b_m obtained from approximate formulas with their exact values, we give an example in which the function is represented analytically.

Example. Consider a periodic function with period 2π:

$$y = f(x) \begin{cases} x/\pi & \text{for } 0 \leqslant x \leqslant \pi,\\ 1 & \text{for } \pi < x < 2\pi,\\ 0 & \text{for } x = 2\pi. \end{cases}$$

We compile a table

x_h	0	$\frac{\pi}{6}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$	$\frac{2\pi}{3}$	$\frac{5\pi}{6}$	π	$\frac{7\pi}{6}$	$\frac{4\pi}{3}$	$\frac{3\pi}{2}$	$\frac{5\pi}{3}$	$\frac{11\pi}{6}$	2π
y_h	0	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{5}{6}$	1	1	1	1	1	1	0

In accordance with Runge's scheme we write the values of y_k and carry out the additions and subtractions indicated in it [see formula (6)]:

$$\begin{array}{lccccccc} & 0 & \frac{1}{6} & \frac{1}{3} & \frac{1}{2} & \frac{2}{3} & \frac{5}{6} & 1\\ & & 1 & 1 & 1 & 1 & 1 & \\ \hline \text{sums} & 0 & \frac{7}{6} & \frac{4}{3} & \frac{3}{2} & \frac{5}{3} & \frac{11}{6} & 1\\ \text{differences} & & -\frac{5}{6} & -\frac{2}{3} & -\frac{1}{2} & -\frac{1}{3} & -\frac{1}{6} & \end{array}$$

Next we perform subtractions and additions with respect to the sums and differences obtained [see formula (7)]:

		sums				differences		
	0	$\frac{7}{6}$	$\frac{4}{3}$	$\frac{3}{2}$		$-\frac{5}{6}$	$-\frac{2}{3}$	$-\frac{1}{2}$
	1	$\frac{11}{6}$	$\frac{5}{3}$			$-\frac{1}{6}$	$-\frac{1}{3}$	
sums	1	3	3	$\frac{3}{2}$	sums	-1	-1	$-\frac{1}{2}$
differences	-1	$-\frac{2}{3}$	$-\frac{1}{3}$		differences	$-\frac{2}{3}$	$-\frac{1}{3}$	

Then we write the expressions for a_m and b_m:

$$6a_0 = 1+3+3+\frac{3}{2}, \qquad 6b_1 = 0.5(-1)+0.866(-1)-\frac{1}{2},$$

$$6a_1 = -1-\frac{2}{3}\cdot 0.866-0.5\cdot\frac{1}{3}, \quad 6b_2 = 0.866\left(-\frac{2}{3}-\frac{1}{3}\right),$$

$$6a_2 = \left(1-\frac{3}{2}\right)+0.5(3-3), \qquad 6b_3 = -1-\left(-\frac{1}{2}\right).$$

$$6a_3 = -1-\left(-\frac{1}{3}\right),$$

Hence

$$a_0 = 1.417,\ a_1 = -0.291,\ a_2 = -0.083,\ a_3 = -0.111,$$
$$b_1 = -0.311;\ b_2 = -0.144,\ b_3 = -0.083.$$

To make the comparison, we give the exact values of the coefficients:

$$a_0 = 1.500,\ a_1 = -0.203,\ a_2 = 0.000,\ a_3 = -0.022,$$
$$b_1 = -0.318,\ b_2 = -0.159,\ b_3 = -0.106.$$

To obtain more accurate values of the coefficients from approximate formulas, we can use schemes with a larger number of ordinates.

Note that the practical harmonic analysis makes it possible to obtain analytic expressions which would approximate the given functions with the least mean square error.

7.19. Backward Interpolation

Interpolation of functions proves to be a useful apparatus in problem solving. The problem of determining a root of an equation or a root of a function is a typical example of the use of an interpolating polynomial.

Consider the following problem of backward interpolation. A function $f(x)$ continuous on the interval $[a, b]$ is specified on the net Λ_{2k}: $a \leqslant \ldots \leqslant x_{-k} < \ldots < x_0 < \ldots < x_k < \ldots \leqslant b$ by its values f_i $(i = 0, \pm 1, \ldots, \pm k)$. We have to find the value of the argument $x^* \in (x_0, x_1)$ corresponding to the specified value of the function $f^* = f_0 + \theta(f_1 - f_0)$, $\theta \in (0, 1)$. The interval (x_0, x_1) is assumed to be so small that x^* is unique.

In essence, we have to find here the root of the equation

$$f(x) = f^*. \tag{1}$$

One of the possible ways of solving this problem is the following. We approximate the function $f(x)$ by its interpolating polynomial $P_n(f, x)$ and replace equation (1) by the equation

$$P_n(f, x) = f^*. \tag{2}$$

We find the real root $\bar{x}^*$ of equation (2) belonging to the interval (x_0, x_1). For all practical purposes we get only an approximate solution of equation (2), i.e. the value $\bar{\bar{x}}^*$. Now we assume that $x^* = \bar{\bar{x}}^*$.

Let us estimate the error of this solution. Assume that the total error of interpolation is Δ, i.e.

$$|P_n(f, x) - f(x)| \leqslant \Delta \tag{3}$$

and the error of the solution of equation (2) is ε, i.e.

$$|\bar{x}^* - \bar{\bar{x}}^*| \leqslant \varepsilon. \tag{4}$$

Then the increment of the function f at the point $\bar{x}^*$ can be represented as

$$f^* - f(\bar{x}^*) - (x^* - \bar{x}^*) f'(\xi), \quad \xi = \bar{x}^* + \theta(x^* - \bar{x}^*), \theta \in (0, 1).$$

From this, with due account of the fact that $f^* = P_n(f, \bar{x}^*)$, we have

$$P_n(f, \bar{x}^*) - f(x^*) = (x^* - \bar{x}^*) f'(\xi).$$

We assume now that $\min\limits_{[x_0, x_1]} |f'(x)| = m_1 > 0$, and, using estimate (3), we obtain

$$|x^* - \bar{x}^*| \leqslant \Delta/m_1. \tag{5}$$

Furthermore,

$$|x^* - \bar{\bar{x}}^*| = |x^* - \bar{x}^* + \bar{x}^* - \bar{\bar{x}}^*| \leqslant |x^* - \bar{x}^*| + |\bar{x}^* - \bar{\bar{x}}^*|$$

and, by virtue of inequalities (4) and (5), we finally get

$$|x^* - \bar{\bar{x}}^*| \leqslant \frac{\Delta}{m_1} + \varepsilon. \qquad (6)$$

Thus, both the solution of the problem posed and error (6) are defined by two processes, the construction of an interpolating polynomial and the solution of equation (2), i.e. the search for the roots of the interpolating polynomial.

These two moments may seem to be unrelated. This is not so, however.

It should be borne in mind that a rise in the degree of the polynomial decreases the error Δ on the one hand and increases the labour needed to solve equation (2) on the other.

Therefore, the degree of the interpolating polynomial must be the lowest in order to achieve the required accuracy.

In the practical solution of a problem of backward interpolation on a uniform net, Stirling's and Bessel's polynomials are usually taken as interpolating polynomials. In that case, equation (2), written with respect to the variable $t = (x - x_0)/h$, is reduced to the form $t = \varphi(t)$ and is solved with the use of an iterative method.

When using Stirling's polynomial, we have

$$t = \frac{1}{\mu f_0^1}\left(f^* - f_0 - \frac{f_0^2}{2!}t^2 - \frac{\mu f_0^3}{3!}t(t^2 - 1^2) - \ldots\right). \qquad (7)$$

The use of Bessel's polynomial gives

$$t = \frac{1}{2} + \frac{1}{f_{1/2}^1}\left(f^* - \mu f_{1/2} - \frac{\mu f_{1/2}^2}{2!}t(t-1)\right.$$
$$\left. - \frac{f_{1/2}^3}{3!}t\left(t - \frac{1}{2}\right)(t-1) - \ldots\right). \qquad (8)$$

As the initial approximation t^0 we take $\frac{1}{\mu f_0^1}(f^* - f_0)$ in the first case and 0.5 or $\frac{1}{2} + \frac{1}{f_{1/2}^1}(f^* - \mu f_{1/2})$ in the second case. When t^*, which is the solution of equation (7) or (8), is obtained, x^* can be found from the formula $x^* = x_0 + t^*h$.

Similarly, if a necessity arises, we can get a solution of the set problem by means of Newton's first or second interpolating polynomial.

Let us consider the second method of solving a problem of backward interpolation based on the existence of a function $g(y)$ which is the inverse of $f(x)$.

Assume that the function $g(y)$ is continuous with a sufficient number of its derivatives on a minimum interval containing the values $y_i = f_i$ $(i = 0, \pm 1, \ldots)$ and $y^* = f^*$. In this case the search for x^* is equivalent to the search for the inverse function $g(y)$ defined by its values x_i at the nodes y_i, at the point $y = f^*$, since $x^* = g(f^*)$.

We have thus reduced the given problem to the problem of interpolating the inverse function $g(y)$ and calculating $g(f^*)$.

This method of solving a problem of backward interpolation is more efficient than the method which includes a solution of an equation as one of its stages. It is especially convenient when we have to find a solution of a problem for a large number of values f^* or to obtain an explicit expression for the root of equation (1). The drawback of the second method is the requirement that a smooth inverse function should exist, a condition which cannot always be fulfilled (for instance this requirement cannot be satisfied for nonmonotonic functions).

It should be pointed out in conclusion that to calculate x^* by means of an inverse function, Aitken's iterative interpolation presented in 7.13 is most convenient.

Example 1. Using the table of values of the function $f = 3^x$ given in the example in 7.9, find out the value of the argument x^* to which the value of the function $f^* = 5$ corresponds. Estimate the error.

△ In the example in 7.9 the order of correctness of the table is 3. Since this value of f^* is at the end of the table, it follows that to calculate x^*, we must use Newton's second interpolating polynomial of the third degree. Setting $x_0 = 1.50$ and $t = (x - x_0)/h$ and using formula (10) from 7.9, we get an equation for determining t^*:

$$5 = 5.196 + \frac{1.248}{1!} t + \frac{0.300}{2!} t(t+1) + \frac{0.072}{3!} t(t+1)(t+2).$$

Taking into account the results of the example from 7.9, we calculate the error of the value of $\bar{x}^*$ from formula (5). Since $\Delta = \Delta_1 + \Delta_2 = 0.004$ and $m_1 = \min\limits_{[0.75,\ 1.5]} | 3^x \ln 3 | \cong 2.5$, the required error constitutes $| x^* - \bar{x}^* | \leqslant \Delta/m_1 = 0.0016$.

We reduce the equation for t^* to a form convenient for the use of an iterative method:

$$t = \frac{5-5.196}{1.248} - \frac{1}{1.248}\left[\frac{0.300}{2!}\,t\,(t+1) + \frac{0.072}{3!}\,t\,(t+1)\,(t+2)\right]$$

and solve it taking $t_0 = (5 - 5.196) \div 1.248 \cong -0.16$ as the initial approximation. Then

$$t_1 = -0.16 - \frac{1}{1.248}\left[\frac{0.300}{2!}\,(-0.16)\,(0.84) + \frac{0.072}{3!}\,(-0.16)\,(0.84)\,(1.84)\right] = -0.141,$$

$$t_2 = -0.16 - \frac{1}{1.248}\left[\frac{0.300}{2!}\,(-0.141)\,(0.859) + \frac{0.072}{3}\,(-0.141)\,(0.859)\,(1.859)\right] = -0.143.$$

We can thus take the value $t^* = -0.14 \pm 0.003$ as the approximation of the solution of the equation. Hence

$$\bar{\bar{x}}^* = x_0 + t^*h = 1.465$$

and the error of the solution of the equation

$$| \bar{x}^* - \bar{\bar{x}}^* | = \varepsilon < 0.0008.$$

Thus the final solution is $x^* = 1.465 \pm 0.003$. ▲

Example 2. Using the table of values of the function $f = \ln x$ given in Example 1 in 7.13 (see p. 340), calculate e^2 with an accuracy of 0.01.

△ The function f has an inverse $g(y) = e^y$ which is continuous together with its derivatives on the interval $(-\infty, \infty)$. We can therefore reduce the calculation of e^y to the calculation, at the point $y = y^* = 2$, of the function e^y given as a table

y	0.00	0.69	1.39	1.61	2.08	2.30
g	1	2	4	5	8	10

We use Aitken's method to solve this interpolation problem. We enumerate the nodes y_i as follows: $y_0 = 2.08$, $y_1 = 2.30$, $y_2 = 1.61$, $y_3 = 1.39$, $y_4 = 0.69$, $y_5 = 0.00$. Using now formula (1) from 7.13 and replacing x by y and x_m by y_m, we calculate the values of the interpolating polynomials $P_n\ (y^*)$:

$$P_1^{01}(2)=\frac{1}{0.22}\cdot\begin{vmatrix}-0.08 & 8\\ -0.30 & 10\end{vmatrix}=7.27,$$

$$P_1^{02}(2)=\frac{1}{0.47}\cdot\begin{vmatrix}0.39 & 5\\ -0.08 & 8\end{vmatrix}=7.49,$$

$$P_2^{012}(2)=\frac{1}{0.69}\cdot\begin{vmatrix}0.39 & 7.49\\ -0.30 & 7.27\end{vmatrix}=7.37,$$

$$P_1^{23}(2)=\frac{1}{0.22}\cdot\begin{vmatrix}0.61 & 4\\ 0.39 & 5\end{vmatrix}=6.77,$$

$$P_2^{023}(2)=\frac{1}{0.69}\cdot\begin{vmatrix}0.61 & 6.77\\ 0.08 & 7.49\end{vmatrix}=7.41,$$

$$P_3^{0123}(x)=\frac{1}{0.91}\cdot\begin{vmatrix}0.61 & 7.41\\ -0.30 & 7.37\end{vmatrix}=7.38.$$

Since $|P_3^{0123}(2)-P_2^{012}(2)|=0.01$ and the required accuracy is attained, we terminate the calculations and set $e^2=7.38\pm 0.01$. ▲

Exercises

1. The function $y=f(x)$ is given as a table

x	1.522	1.523	1.524
y	20.477	20.906	21.354

Find its value at the point $x=1.5228$ using Newton's first interpolation formula.

2. The function $y=f(x)$ is given as a table

x	1.529	1.530	1.531
y	23.911	24.498	25.115

Find its value at the point $x=1.5303$ using Newton's second interpolation formula.

3. Construct Lagrange's interpolating polynomial for the function given as a table

x	−2	−1	2	3
y	−12	−8	3	5

4. Construct Lagrange's interpolating polynomial for the function $f(x)=e^{-x}$ if the points $x_0=1$, $x_1=2$, $x_2=3$ are interpolation nodes. Estimate the error for $x=1.5$.

5. Compile a table of finite differences for the function given as a table

x	−2	−1	0	1	2	3
y	10	5	1	−15	−20	−100

6. Compile a table of divided differences for the function given as a table

x	-3	1	0	2	3
y	-15	-7	1	25	47

7. For the function $y = f(x)$ given as a table

x	1.03	1.08	1.016	1.23	1.26	1.33	1.39
y	2.80107	2.94468	3.18993	3.42123	3.52542	3.78104	4.01485

calculate the value at the point $x = 1.21555$ with an accuracy of 10^{-5} using Aitken's method.

8. Using Stirling's formula, find the value of the function $y = f(x)$ at the point $x = 1.34627$ if the function is given as a table

x	1.335	1.340	1.345	1.350	1.355	1.360
y	4.16206	4.25562	4.35325	4.45522	4.56184	4.67433

9. For the function given as a table

x	1.435	1.440	1.445
y	0.892687	0.893698	0.894700

determine the value of the argument corresponding to the value of the function 0.892914.

10. Use the method of backward interpolation to find, with an accuracy of 10^{-5}, the root of the equation $\sqrt{x+1} - \frac{1}{x} = 0$ which lies on the interval [0.7, 0.8].

11. Construct an interpolating polynomial for a function given as a table

x	1	2	3
y	384.6	507.9	477.9

12. Construct an interpolating polynomial for a function given as a table

x	1	2	3
y	349.1	416.9	430.6

13. Calculate the value of the function $f(x)$ at the point $x = x_1$ using a requisite interpolating polynomial and employing four-digit tables of trigonometric functions with the stepsize of 1°. Estimate the absolute error of the result in the following cases: (a) $f(x) = \sin x$, $x_1 = 37.7°$, (b) $f(x) = \cos x$, $x_1 = 19°48'$,

(c) $f(x) = \sin x$, $x_1 = 53°12'$, (d) $f(x) = \cos x$, $x_1 = 36°48'$, (e) $f(x) = \cos x$, $x_1 = 71°6'$, (f) $f(x) = \tan x$, $x_1 = 67°48'$.

14. Set up a trigonometric interpolating polynomial for the function $y = f(x)$ defined in the interval $(0, \pi)$ and given as a table of values of $y_k = f(x_k)$:

k	0	1	2	3
x_k	0	$\pi/4$	$\pi/2$	$3\pi/4$
y_k	1	2	2.4	2.6

15. Set up a trigonometric polynomial of the order not lower than the second for the function $y = f(x)$ defined in the interval $(0, 1)$ and given as a table for the values of $y_k = f(x_k)$:

k	0	1	2	3	4	5
x_k	0	1/6	1/3	1/2	2/3	5/6
y_k	1	0	−2	−3	0	2

Chapter 8

Numerical Differentiation and Integration

8.1. Statement of a Problem and the Basic Formulas for Numerical Differentiation

When solving practical problems, it is often necessary to obtain the values of the derivatives of various orders of the function f given as a table or as a complicated analytic expression. In that case a direct use of the methods of differential calculus is either impossible or difficult. Then use is made of approximate methods of numerical differentiation.

The simplest expressions for derivatives result from differentiation of interpolation formulas.

Consider the following problem. The values f_i of the function f, continuously differentiable $n + 1 + m$ times, are defined on the net $a \leqslant x_0 < x_1 < \ldots < x_n \leqslant b$ at the nodal points x_i. We have to find the derivative $f^{(m)}(x^*)$, $x^* \in [a, b]$ and estimate the error.

One of the possible methods of solving this problem is the following. Using the nodal points x_i $(i = 0, 1, \ldots, n)$, we construct for the function f an interpolating polynomial with the remainder R_n such that

$$f(x) = P_n(x) + R_n(x). \tag{1}$$

We differentiate the right-hand and left-hand sides of relation (1) m times and set $x = x^*$:

$$f^{(m)}(x^*) = P_n^{(m)}(x^*) + R_n^{(m)}(x^*). \tag{2}$$

For sufficiently smooth functions, i.e. for functions with bounded derivatives, sufficient number of nodes and sufficient accuracy of calculations, the quantity $R_n^{(m)}(x^*)$ is small and $P_n^{(m)}(x^*)$ is a good approximation for $f^{(m)}(x^*)$, and so we can set

$$f^{(m)}(x^*) \cong P_n^{(m)}(x^*). \tag{3}$$

In practical computations, numerical differentiation proves to be very sensitive to errors in the initial data, to the discarding of terms of a series and other operations of this kind. In addition, the high accuracy of interpolation the [smallness of $R_n(x)$], does not guarantee a high accuracy of the interpolation formula for the derivatives [the smallness of $R_n^{(m)}(x)$]. Therefore one must be careful in applying the methods of numerical differentiation, using it, as a rule, for small m.

Bearing in mind all we have said and also the fact that the calculation of derivatives of higher orders can be reduced to a successive calculation of lower-order derivatives, we shall consider in more detail the technique of obtaining formulas for computing f' and f'' at the nodal points of a uniform net. To obtain derivatives at the nodal points, it is expedient to use Stirling's interpolating polynomial and its remainder [see formulas (5) and (6) in 7.8]. Thus, differentiating Stirling's polynomial and its remainder with respect to x and setting $x^* = x_0$ $(t^* = 0)$, we get the following expressions for the derivative:

$$f'(x_0) = \frac{1}{h}\mu f_0^1 \pm \frac{M_3}{6}h^2 \quad (k=1), \tag{4}$$

$$f'(x_0) = \frac{1}{h}\left(\mu f_0^1 - \frac{\mu f_0^3}{6}\right) \pm \frac{M_5}{30}h^4 \quad (k=2). \tag{5}$$

Differentiating Stirling's polynomial twice with respect to x and calculating the value of the second derivative at the point $x^* = x_0$, we have

$$f''(x_0) = \frac{1}{h^2} f_0^2 \pm \frac{M_4}{12}h^2 \quad (k=1), \tag{6}$$

$$f''(x_0) = \frac{1}{h^2}\left(f_0^2 - \frac{1}{12} f_0^4\right) \pm \frac{M_6}{90}h^4 \quad (k=2). \tag{7}$$

The derivative at the exact middle point between the nodes $x^* = x_0 + \frac{h}{2}$ can be calculated with the use of Bessel's polynomial. In this case the appropriate formulas for the derivative have the form

$$f'\left(x_0 + \frac{h}{2}\right) = \frac{1}{h} f_{1/2}^1 \pm \frac{M_3}{24}h^2 \quad (k=1), \tag{8}$$

$$f'\left(x_0 + \frac{h}{2}\right) = \frac{1}{h}\left(f_{1/2}^1 - \frac{1}{24} f_{1/2}^3\right) \pm \frac{3M_5}{640}h^4 \quad (k=2). \tag{9}$$

Of practical interest are the so-called formulas for one-sided differentiation which make it possible to calculate $f'(x_0)$ using the nodal points $x_i = x_0 + ih$ $(i = 0, 1, \ldots, k, \ldots$ or $i = 0, -1, \ldots, -k, \ldots)$. It is convenient to set up these formulas with the use of Newton's first and second interpolating polynomials.

Differentiating Newton's first polynomial with respect to x and calculating the value of the derivative at the point $x = x_0$ $(t = 0)$ for $k = 1$ and $k = 2$, we get the formulas

$$f'(x_0) = \frac{1}{h}\Delta f_0 \pm \frac{1}{2}M_2 h, \tag{10}$$

$$f'(x_0) = \frac{1}{h}\left(\Delta f_0 - \frac{1}{2}\Delta^2 f_0\right) \pm \frac{1}{3}M_3 h^2 \tag{11}$$

respectively.

Similarly, differentiating Newton's second polynomial for $k = -1$ and $k = -2$ we obtain

$$f'(x_0) = \frac{1}{h}\nabla f_0 \pm \frac{1}{2}M_2 h, \tag{12}$$

$$f'(x_0) = \frac{1}{h}\left(\nabla f_0 + \frac{1}{2}\nabla^2 f_0\right) \pm \frac{1}{3}M_3 h^2 \tag{13}$$

respectively.

8.2. Peculiarities of Numerical Differentiation

Here we give again all the second-order formulas expressing finite differences appearing in them directly in terms of the values f_i of the function. From relations (4), (6) and (8) from 8.1 we have

$$f'(x_0) = \frac{f_1 - f_{-1}}{2h} \pm \frac{M_3}{6}h^2, \tag{1}$$

$$f''(x_0) = \frac{f_1 - 2f_0 + f_{-1}}{h^2} \pm \frac{M_4}{12}h^2, \tag{2}$$

$$f'\left(x_0 + \frac{h}{2}\right) = \frac{f_1 - f_0}{h} \pm \frac{M_3}{24}h^2. \tag{3}$$

Relations (11) and (13) from 8.1 yield, respectively, formulas

$$f'(x_0) = \frac{1}{2h}(-3f_0 + 4f_1 - f_2) \pm \frac{M_3}{3}h^2, \tag{4}$$

$$f'(x_0) = \frac{1}{2h}(3f_0 - 4f_{-1} + f_{-2}) \pm \frac{M_3}{3}h^2. \tag{5}$$

We can see from the formulas given above that the error of the method diminishes with a decrease in the stepsize of the net. However, if the values f_i of the function are given approximately, say, with the same absolute error ε, then the total error of the formulas for numerical differentiation will include an additional term which is inversely proportional to h^m (m is the order of the derivative). Therefore, it is reasonable to decrease h only to a certain limit.

To illustrate the aforesaid, we shall consider the right-hand side of formula (3). Its total error constitutes

$$\Delta = \frac{M_3}{24} h^2 + \frac{2\varepsilon}{h}. \tag{6}$$

Equating $\Delta'(h)$ to zero, we get the point of extremum of the function $\Delta(h)$:

$$h_0 = 2\sqrt[3]{\frac{3\varepsilon}{M_3}} \cong 2.9\sqrt[3]{\frac{\varepsilon}{M_3}}. \tag{7}$$

Since $\Delta''(h) > 0$, it follows that h_0 is the point of minimum of $\Delta(h)$ and

$$\Delta(h_0) = \frac{3}{2}\sqrt[3]{\frac{1}{3} M_3 \varepsilon^2} \cong \sqrt[3]{M_3 \varepsilon^2}. \tag{8}$$

This relation means, in particular, that we cannot guarantee, for any h, that the error of the result will be the quantity $o(\varepsilon^{2/3})$.

Similarly, using formula (2) for the optimum stepsize, we get an expression

$$h_0 = 2\sqrt[4]{\frac{3\varepsilon}{M_4}} \cong 2.6\sqrt[4]{\frac{\varepsilon}{M_4}}, \tag{9}$$

and, using formulas (4) and (5), we get an expression

$$h_0 = \sqrt[3]{\frac{6\varepsilon}{M_3}} \cong 1.8\sqrt[3]{\frac{\varepsilon}{M_3}}. \tag{10}$$

Thus, when calculating derivatives, we must first find the optimum stepsize of the initial table of values f_i.

Example 1. Calculate $f'(1.6)$ and $f''(1.4)$ for the function $f = \ln x$ given as a table

x	1.2	1.3	1.4	1.5	1.6
f	0.1823	0.2626	0.3364	0.4054	0.4700

which contains the values f_i with all valid values in the broad sense. Evaluate the error of the result.

△ To calculate the required derivatives, we apply formulas (5) and (2) respectively. Then, using relations (10) and (9) and the initial data, we get the following values for the optimum stepsize:

$$h_{01}=\sqrt[3]{\frac{6\cdot10^{-4}}{0.73}}\cong 0.1 \text{ when calculating } f'(1.6)$$

$$h_{02}=\sqrt[4]{\frac{48\cdot10^{-4}}{2.1}}\cong 0.22 \text{ when calculating } f''(1.4).$$

Since the tabular data do not allow us to choose 0.22 as the stepsize, we take the closest possible number, 0.2, as h_2. Consequently,

$$f'(1.6)=\frac{1}{0.2}(3\cdot0.4700-4\cdot0.4054+0.3364)=0.624,$$

the total error not exceeding

$$\Delta=\frac{0.73}{3}\cdot0.1^2+\frac{4\cdot10^{-4}}{0.1}=0.007,$$

and $$f''(1.4)=\frac{1}{0.2^2}(0.4700-2\cdot0.3364+0.1823)=-0.512,$$

the total error not exceeding

$$\Delta=\frac{2.9}{12}\cdot0.2^2+\frac{4\cdot10^{-4}}{0.2^2}=0.02.$$

Although the estimates of the error are as a rule, too high, still it shows that the operation of finding the second derivative is more reliable than that of finding the first derivative. ▲

In some practical cases we have to find the derivative being given only a table of values of the function. Then it is evidently impossible to evaluate the error. We calculate the approximate values of the derivatives directly from one of the formulas (4)-(13) from 8.1, disregarding the error.

Example 2. Calculate $f'(1.3)$, $f''(1.4)$ for the function $f(x)$ given as a table

x	1.2	1.3	1.4	1.5	1.6
$y=f(x)$	0.18	0.26	0.34	0.41	0.47

△ From formulas (4) and (6) given in 8.1 we obtain, respectively,

$$f'(1.3)=\frac{1}{0.1}\cdot\frac{1}{2}(0.34-0.26+0.26-0.18)=0.8,$$

$$f''(1.4)=\frac{1}{0.1^2}(0.41-0.34-0.34+0.26)=-1. \ ▲$$

8.3. Statement of a Problem of Numerical Integration

Assume that we have to calculate the integral

$$I=\int_a^b f(x)\,\mathrm{d}x. \tag{1}$$

We know from analysis that for the function f continuous on the interval $[a, b]$ integral (1) exists and is equal to the difference of the values of the antiderivative F of the function f at the points b and a:

$$I=\int_a^b f(x)\,\mathrm{d}x=F(b)-F(a). \tag{2}$$

However, in the majority of practical problems it is impossible to express the antiderivative in terms of elementary functions. In addition, the function f is often given as a table of its values for definite values of the argument. All this makes it necessary to use approximate methods of calculation of integral (1) which can be conventionally divided into analytical and numerical methods. The former consist, in essence, in the construction of an antiderivative and further use of formula (2). As to the latter, they make it possible to find the numerical value of the integral from the known values of the integrand function (and sometimes of its derivatives) at given points known as *nodes*. In this chapter we consider only numerical methods of integrating a function. The process of numerical calculation of an integral is known as *quadrature* and the corresponding formulas as *quadrature formulas*.

Depending on the method of definition of an integrand function, we shall consider two distinct, in the sense of their realization, cases of numerical integration.

Problem I. The values f_i of a function f which belongs to a definite class F are given on the interval $[a, b]$ at the nodes x_i. We have to approximate integral (1) and estimate the error of the value obtained.

This is the usual statement of the problem of numerical integration when the integrand function is given as a table.

Problem II. On the interval $[a, b]$ the function $f(x)$ is given in the form of an analytical expression. We have to calculate integral (1) with a specified limiting error ε.

One of the possible methods of solving the problems stated is based on the use of various quadrature formulas of the form

$$I \equiv \int_a^b f(x)\,dx \cong (b-a)\sum_{i=1}^{n} A_i f(x_i) \equiv I_n \qquad (3)$$

with the known remainder $R_n[f] = I - I_n$ or its estimate.

In the general case the nodes x_i and the weight factors (weights) A_i are not known in advance and must be found when each quadrature formula (3) is derived, on the basis of the requirements imposed on it.

In essence, the problem of numerical integration is equivalent to the estimation of the mean value of a function. Indeed, the mean value of a function on the interval $[a, b]$ is determined as follows:

$$\bar{f} = \frac{1}{b-a}\int_a^b f(x)\,dx$$

and therefore

$$\int_a^b f(x)\,dx = (b-a)\bar{f}.$$

In its turn, the calculation of the mean value of a function is a statistical problem which includes the problems of successive sampling and planning an experiment. Since it is difficult to pose such a problem, we shall consider in this chapter only classical methods of numerical integration based on the preliminary definition of the nodes at which the information on the function being integrated must be given and of the information itself.

We pass now to the algorithms of the solution of the problems formulated above.

Algorithm of solution of Problem I.

1°. We choose a definite quadrature formula (3) and calculate I_n. If the values f_i of the function are approximate, then we calculate, in fact, only the approximate value $\overline{I}_n$ for the exact value I_n.

2°. We assume approximately that $I \cong \overline{I}_n$.

3°. Using a definite expression for the remainder or its estimate for the chosen quadrature formula, we calculate the error of the method:

$$\Delta_1 = | I - I_n | = | R_n$$

4°. We calculate the computing error $\overline{I}_n$:

$$\Delta_2 = | I_n - \overline{I}_n |$$

using the errors of the approximate values f_i.

5°. We find the total absolute error of the approximate value I_n.

$$\Lambda = | I - \overline{I}_n | \leqslant \Lambda_1 + \Lambda_2.$$

6°. We obtain the solution of the problem in the form

$$I = \overline{I}_n \pm \Delta.$$

For sufficiently smooth functions, i.e. for functions with a limited variation of the derivatives, the error of the quadrature formulas (3), for sufficiently large n, is small as a rule. Therefore, if the initial values f_i and the calculations of $\overline{I}_n$ are sufficiently accurate, we can expect that $\overline{I}_n$ will be a close approximation of I. These considerations serve as the basis for the following algorithm.

Algorithm of solution of Problem II.

1°. We represent ε as the sum of three nonnegative terms:

$$\varepsilon = \varepsilon_1 + \varepsilon_2 + \varepsilon_3,$$

where ε_1 is the limiting error of the method, ε_2 is the limiting error of computation of $\overline{I}_n$, ε_3 is the limiting error of rounding off the result.

2°. We choose n in the quadrature formula such that the inequality

$$\Lambda_1 = | I - I_n | = | R_n | \leqslant \varepsilon_1$$

is satisfied.

3°. We calculate f_i with an accuracy which would ensure the validity of the inequality

$$\Delta_2 = | I_n - \bar{I}_n | \leqslant \varepsilon_2$$

when I_n is calculated from formula (3). To do this, it is evidently sufficient to calculate all f_i with the absolute error $\dfrac{\varepsilon}{(b-a)\sum\limits_{i=1}^{n} |A_i|}$.

4° We round off the value of $\bar{I}_n$ found in item 3° (if $\varepsilon_3 \neq 0$) with the limiting error ε_3 to the value $\bar{\bar{I}}_n$.

5°. We obtain the solution of the problem in the form

$$I = \bar{\bar{I}}_n \pm \varepsilon.$$

As we have mentioned, the quadrature formulas used in the algorithms of the two problems are constructed on the basis of some criteria defining the position of the nodes and the values of the weight factors. Here are some of the criteria that may be used: representation of an integral as an integral sum, approximation of an integrand function (say, by a polynomial) followed by the integration of the approximating function, the requirement that formula (3) should be absolutely exact for a definite class of functions, and others.

8.4. Basic Quadrature Formulas

Rectangular formulas. As is known, by virtue of its construction the definite integral is the limit of integral sums

$$\int_a^b f(x)\,dx = \lim_{\max h_i \to 0} \sum_{i=1}^{n} h_i f(\xi_i), \tag{1}$$

each of which corresponds to a certain division D_n: $a = x_0 < x_1 < \ldots < x_n = b$ of the interval $[a, b]$ and to the arbitrary collection of points $\xi_i \in [x_{i-1}, x_i]$ for each division, $h_i = x_i - x_{i-1}$.

Restricting our consideration to a finite number of terms on the right-hand side of relation (1) and taking particular values of the argument belonging to the interval $[x_{i-1}, x_i]$ as a collection of points ξ_i, we can get various formulas for approximate integration. Thus, taking the values of the left-hand and right-hand endpoints of the interval $[x_{i-1}, x_i]$ as ξ_i, we get a *formula of the left or right rectangles* respectively $(h_i = 1/n = \text{const})$:

$$I \equiv \int_a^b f(x)\,\mathrm{d}x \cong (b-a) \sum_{i=0}^{n-1} \frac{1}{n} f_i \equiv I_1, \tag{2}$$

$$I \equiv \int_a^b f(x)\,\mathrm{d}x \cong (b-a) \sum_{i=1}^{n} \frac{1}{n} f_i \equiv I_r. \tag{3}$$

The names of these formulas are due to their geometrical interpretation. If we construct a curve $y = f(x)$ in

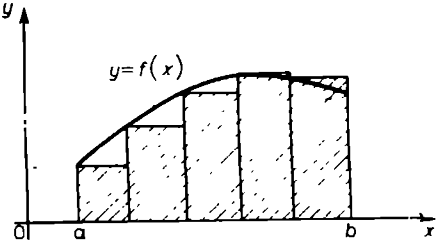


Fig. 8.1

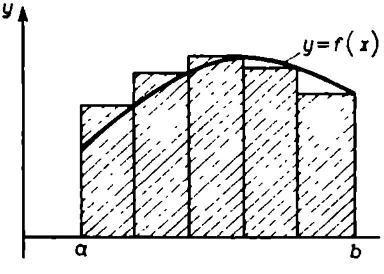


Fig. 8.2

the xy-plane and divide the interval $[a, b]$ into n parts by the points x_i of the net D_n, then the formula of the left rectangles yields, as an approximate value of the integral, the total area of the hatched rectangles shown in Fig. 8.1 and the formula of the right rectangles yields the total area of the hatched rectangles shown in Fig. 8.2.

Example 1. Using the formulas of the left and right rectangles, calculate $\int\limits_1^9 \frac{dx}{x+2}$ setting $n=4$.

△ Knowing the integration limits $a=1$ and $b=9$, we find the step $h=(b-a)/n=2$. Then $x_0=1$, $x_1=3$, $x_2=5$, $x_3=7$ and $x_4=9$ are the division points and the values of the integrand function $f(x)=1/(x+2)$ at these points are

$$y_0=f(x_0)=1/3,\quad y_1=f(x_1)=1/5,$$
$$y_2=f(x_2)=1/7,\quad y_3=f(x_3)=1/9,\quad y_4=f(x_4)=1/11.$$

Next we find the numerical value of the integral using formula (2):

$$I_l=\frac{b-a}{n}(y_0+y_1+y_2+y_3)=2\left(\frac{1}{3}+\frac{1}{5}+\frac{1}{7}+\frac{1}{9}\right)\cong 1.6024.$$

Now if we use formula (3) to calculate the definite integral, then we obtain

$$I_r=\frac{b-a}{n}(y_1+y_2+y_3+y_4)=2\left(\frac{1}{5}+\frac{1}{7}+\frac{1}{9}+\frac{1}{11}\right)\cong 1.1053.\ ▲$$

The *rectangular formula*, where the midpoints of the intervals $[x_{i-1}, x_i]$ are taken as ξ_i, is the most widely used formula based on the representation of a definite integral as an integral sum. For a uniform net $(h_i=h)$ this formula has the form

$$I\equiv\int\limits_a^b f(x)\,dx\cong\frac{b-a}{n}\sum_{i=1}^{n} f_{i-1/2}\equiv I_n, \tag{4}$$

where $f_{i-1/2}=f\left(x_i-\frac{h}{2}\right)$, $x_0=a$, $x_n=b$.

We seek an expression for the remainder of the approximate formula (4). For our purpose, we represent the integral appearing on the left-hand side of relation (4) as a sum

$$\int\limits_a^b f(x)\,dx=\sum_{i=1}^{n}\int\limits_{x_{i-1}}^{x_i} f(x)\,dx. \tag{5}$$

Assuming the function $f(x)$ to be twice differentiable, i.e. $f\in C^2[a, b]$, we write, for the function $f(x)$, on each of the intervals $[x_{i-1}, x_i]$, Taylor's formula with the remainder in Lagrange's form:

$$f(x) = f_{i-1/2} + \left(x - x_i + \frac{h}{2}\right) f'_{i-1/2} + \frac{\left(x - x_i + \frac{h}{2}\right)^2}{2} f''(\eta_i),$$

$$\eta_i \in (x_{i-1}, x_i). \tag{6}$$

Replacing the function f on the right-hand side of relation (5) by its representation (6), we carry out the integration using the second theorem of the mean:

$$\int_a^b f(x)\,dx = (b-a)\frac{1}{n}\sum_{i=1}^{n} f_{i-1/2} + \frac{h^3}{24}\sum_{i=1}^{n} f''(\bar{\eta}_i),$$

$$\bar{\eta}_i \in (x_{i-1}, x_i).$$

Since the second derivative is continuous, there is a point $\eta \in (a, b)$ such that

$$\sum_{i=1}^{n} f''(\bar{\eta}_i) = nf''(\eta) = \frac{b-a}{h} f''(\eta).$$

Using this relation, we finally obtain

$$\int_a^b f(x)\,dx = (b-a)\sum_{i=1}^{n} \frac{1}{n}\cdot f_{i-1/2} + \frac{b-a}{24} h^2 f''(\eta). \tag{7}$$

Comparing formulas (4) and (7), we get an expression for the remainder in the quadrature formula (4):

$$R_n[f] = I - I_n = \frac{b-a}{24} h^2 f''(\eta). \tag{8}$$

We can thus represent the estimate of the error of the quadrature formula (4) in the form

$$\Delta_1 = \left| \int_a^b f(x)\,dx - (b-a)\sum_{i=1}^{n} \frac{1}{n} f_{i-1/2} \right| \leqslant \frac{b-a}{24} h^2 M_2, \tag{9}$$

where $M_2 = \max\limits_{[a,\,b]} |f''(x)|$.

The expressions for the remainder (8) and for the error (9) show that formula (4) is exact for any linear function since the second derivative of such a function is zero and, consequently, the remainder and the error are zero too.

We shall prove that the estimate obtained cannot be improved, i.e. that there is a function for which the error of computing an integral from formula (4) is exactly equal to the right-hand side of (9). For this purpose, we shall consider $f = x^2$ as the function being integrated and apply formula (4) to it:

$$I_n = (b-a)\frac{1}{n}\sum_{i=1}^{n}\left(a+\frac{2i-1}{2}h\right)^2.$$

Removing the brackets under the sign of the sum and carrying out the summation, we obtain

$$I_n = \frac{b^3-a^3}{3} - (b-a)\frac{h^2}{12}.$$

On the other hand, a direct integration of the function x^2 yields

$$I = \int_a^b x^2\,dx = \frac{b^3-a^3}{3}.$$

Setting up a difference of the exact value of the integral and its approximate value, we get the following expression for the remainder:

$$I - I_n = (b-a)\frac{h^2}{12}.$$

Returning to the estimate of error (9) and noting that for the function x^2 the second derivative (and, consequently, M_2) is equal to 2, we get the same value for the error

$$\Delta_1 = (b-a)\frac{h^2}{12},$$

i.e. the estimate of error (9) is attained on the parabola $y = x^2$. This result can be extended to an arbitrary parabola since the operation of integration is linear and formula (4) is exact for linear functions.

Estimate (9) evidently does not take into account the errors of the calculation of I_n. The error Δ_1 reflects the difference between the exact formula of Newton-Leibniz and the approximate formula (4), i.e. is an error of the method.

Let us now estimate the error of the approximate value $\bar{I}_n$. If the values of the function used in the quadrature formula have been obtained by an approximate method or, for some reason, the calculations cannot be absolutely accurate, then a computing error and rounding errors occur. Assume, for instance, that the values $f_{i-1/2}$ in formula (4) have been calculated with the same absolute error ε. Then the total computing error $\bar{I}_n$ constitutes

$$\Delta_2 = (b-a)\sum_{i=1}^{n}\frac{1}{n}\varepsilon = (b-a)\varepsilon. \tag{10}$$

Note a characteristic peculiarity of this error: it does not depend on the number of divisions of the integration interval but is only proportional to its length.

Example 2. Use the rectangular formula to calculate $\int\limits_0^1 \frac{dx}{1+x}$ setting $n = 4$. Estimate the error of the approximate value obtained.

$\triangle$ From the given integration limits and the number of divisions n we find the step: $h = (1-0)/4 = 0.25$. Next, from formula (4) we get

$$I_4 = 0.25\left[f\left(\frac{1}{8}\right)+f\left(\frac{3}{8}\right)+f\left(\frac{5}{8}\right)+f\left(\frac{7}{8}\right)\right].$$

Calculating the necessary values of the function with three valid, in the narrow sense, digits ($\varepsilon = 0.0005$), we obtain

$$\int\limits_0^1 \frac{dx}{1+x} = 0.25\,(0.889+0.727+0.615+0.533) = 0.691.$$

We use formula (9) to estimate the error of the method, for which purpose we first find the maximum of the absolute value of the second derivative of the integrand function:

$$M_2 = \max_{[0,\,1]}\left|\left(\frac{1}{1+x}\right)''\right| = \max_{[0,\,1]}\frac{2}{(1+x)^3} = 2.$$

Thus the error of the method is

$$\Delta_1 \leqslant (1/24)\cdot 0.25^2\cdot 2 \cong 0.0053.$$

Using formula (10), we find the computing error

$$\Delta_2 \leqslant 1\cdot 0.0005 = 0.0005.$$

Consequently, we can take $\Delta = \Delta_1 + \Delta_2 = 0.006$ as the total error of the approximate value of the integral and write the final answer $\int\limits_0^1 \frac{dx}{1+x} = 0.691 \pm 0.006$.

For the sake of comparison we give several digits of the exact value of the integral calculated: $\ln 2 = 0.693147 \ldots$. ▲

Example 3. Use the rectangular formula to calculate the integral $\int\limits_0^1 \frac{dx}{1+x}$ with an accuracy of 0.001.

△ Applying the algorithm of the solution of Problem II from 8.3, we represent the total error as the sum of three terms: $0.001 = 0.0009 + 0.00005 + 0.00005$. Next we choose n from the condition

$$\Delta_1 = \frac{b-a}{24} h^2 M_2 = \frac{b-a}{24} \left(\frac{b-a}{n}\right)^2 M_2 \leqslant 0.0009.$$

Solving this inequality with respect to n, for $b - a = 1$ and $M_2 = 2$, we get $n \geqslant 10$.

We tabulate the values of the function $1/(x + 1)$ with four valid digits in the narrow sense:

0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
0.9254	0.8696	0.8000	0.7407	0.6897	0.6452	0.6061	0.5714	0.5405	0.5128

Using the rectangular formula (4), we obtain

$$\bar{I}_{10} = 0.1\,(0.9524 + 0.8696 + 0.8000 + 0.7407 + 0.6897 + 0.6452 + 0.6061 + 0.5714 + 0.5405 + 0.5128) = 0.69284.$$

Rounding off the result obtained, we have $I = 0.6928 \pm 0.001$. ▲

The trapezoid formula. Let us consider now another method of constructing quadrature formulas connected with the approximation of an integrand function by an interpolating polynomial. We consider a simple case of approximation by a first-order polynomial with nodes at the points a and b:

$$f(x) = f(a) + \frac{x-a}{b-a}\,[f(b) - f(a)] + (x-a)(x-b)\,\frac{f''(\bar{\eta})}{2},$$

$$\bar{\eta} \in (a, b).$$

Integrating the right-hand and left-hand sides of this relation and employing the second theorem of the mean to integrate the last term on the right-hand side, we obtain

$$\int_a^b f(x)\,dx = \frac{b-a}{2}[f(a)+f(b)] - \frac{(b-a)^3}{12} f''(\eta), \quad \eta \in (a, b).$$

Thus, assuming that the interval of integration is small, we get a quadrature formula known as the *trapezoid formula*:

$$I = \int_a^b f(x)\,dx \cong \frac{b-a}{2}[f(a)+f(b)] \equiv I_2 \qquad (11)$$

with a remainder

$$R_2[f] = I - I_2 = -\frac{(b-a)^3}{12} f''(\eta), \quad \eta \in (a, b). \qquad (12)$$

Using expression (12) for the remainder, we can represent the estimate of the error of the quadrature formula (11) as

$$\Delta_1 = \left| \int_a^b f(x)\,dx - \frac{b-a}{2}(f(a)+f(b)) \right| \leqslant \frac{(b-a)^3}{12} M_2, \qquad (13)$$

where $M_2 = \max\limits_{[a,b]} |f''(x)|$.

The expressions obtained for remainder (12) and error (13) show that the quadrature formula (11) is exact for all linear functions since the second derivative of functions of this kind is zero and, consequently, the remainder and the error are also zero.

By analogy with what we have done for estimate (9), we can show that estimate (13) cannot be improved since it is attained on an arbitrary parabola.

When formula (11) is used to estimate the computing error for the case, when the values of the function have been calculated with the same accuracy ε, the estimate has the form

$$\Delta_2 \leqslant \frac{b-a}{2}(\varepsilon + \varepsilon) = (b-a)\,\varepsilon. \qquad (14)$$

Note that the computing errors of the quadrature formulas (11) and (4) are the same.

Example 4. Use the trapezoid formula to calculate the integral

$\int\limits_0^1 \frac{dx}{1+x}$. Estimate the error of the approximate value obtained.

△ From formula (11) we have

$$I_2 = 0.5\,[f(0) + f(1)].$$

Calculating the required values of the function, we get

$$\int\limits_0^1 \frac{dx}{1+x} \cong 0.5\,(1+0.5) = 0.75.$$

We find the error of the method from formula (13) using the value $M = 2$ obtained in Example 2:

$$\Delta_1 \leqslant \frac{1^3}{12}\cdot 2 \cong 0.17.$$

The computing error is evidently zero since the values of the function and I_2 have been found with the absolute accuracy.

Thus the final result is $\int\limits_0^1 \frac{dx}{1+x} = 0.75 + 0.17$. ▲

Note that in Example 4 we had a considerably less accurate solution than in Example 2. However, we must not draw a hasty conclusion since the use of the trapezoid formula in Example 4 has some advantages. First, whereas the integrand function is given in the form of a table of its values at the nodes x_i, the use of the rectangular formula requires that the values of the function should be also found at the points $x_i \pm h/2$, and this involves additional difficulties and errors. Second, in Example 4 the values of the integrand function were calculated at two points whereas in Example 2 this was done at four points, which naturally took more time.

These arguments show that the importance of the quadrature formula is defined not only by the form of its remainder (the error) but also by other factors, the time taken by calculations, for one.

Other kinds of the quadrature formula. Let us consider one more way of constructing quadrature formulas, that of representing the integral as a linear combination of the values of the integrand function and its derivatives at some nodes x_i followed by determination of the unknown coefficients (weight factors).

Assume, for instance, that we are constructing a quadrature formula of the following kind:

$$\int_a^b f(x)\,dx = (b-a)\,[A_1 f(a) + A_2 f(b) + A_3 f'(a) + A_4 f'(b)]. \tag{15}$$

We find the weight factors A_i ($i = 1, 2, 3, 4$) such that formula (15) is exact for arbitrary polynomials of degree zero, one, two and three. Since the operations of integration and differentiation are linear, this condition is fulfilled if it is fulfilled for the polynomials 1, x, x^2 and x^3.

Substituting these polynomials for $f(x)$ in relation (15), under the condition of its exact validity, we obtain the following system of linear equations for A_i:

$$\begin{cases} A_1 + A_2 = 1, \\ aA_1 + bA_2 + A_3 + A_4 = \frac{1}{2}(a+b), \\ a^2A_1 + b^2A_2 + 2aA_3 + 2bA_4 = \frac{1}{3}(a^2 + ab + b^2), \\ a^3A_1 + b^3A_2 + 3a^2A_3 + 3b^2A_4 = \frac{1}{4}(a^3 + a^2b + ab^2 + b^3). \end{cases}$$

Solving this system, we find that

$$A_1 = A_2 = 1/2,\ A_3 = -A_4 = (b-a)/12.$$

Thus the required quadrature formula has the form

$$I \equiv \int_a^b f(x)\,dx \cong (b-a)\left[\frac{f(a)+f(b)}{2} + (b-a)\,\frac{f'(a)-f'(b)}{12}\right] \equiv I_4. \tag{16}$$

We seek an expression for the remainder of this formula, for which purpose we represent the function being integrated as the sum of a third-degree Hermite interpolating polynomial with two double nodes a and b and the remainder and then integrate the right-hand and left-hand sides of that representation on the interval $[a, b]$:

$$\int_a^b f(x)\,dx = \int_a^b H_3(x)\,dx + \int_a^b \frac{(x-a)^2(x-b)^2}{4!} f^{\mathrm{IV}}(\bar{\eta})\,dx,$$

$$\bar{\eta} \in (a, b).$$

The first term on the right-hand side yields the right-hand side of the quadrature formula (16) since this formula is accurate for all third-degree polynomials and, consequently, for the Hermite polynomial $H_3(x)$ as well. The second term on the right-hand side yields an expression for the remainder of formula (16). Using the second mean-value theorem and carrying out the integration, we obtain

$$R_4[f] = I - I_4 = \frac{(b-a)^5}{720} f^{\mathrm{IV}}(\eta), \quad \eta \in (a, b). \tag{17}$$

The expression for the remainder we have obtained makes it possible to write the estimate of the error of the quadrature formula (16) in the form

$$\Delta_1 = |I - I_4| \leqslant \frac{(b-a)^5}{720} M_4, \tag{18}$$

where $M_4 = \max\limits_{[a,\,b]} |f^{\mathrm{IV}}(x)|$.

Estimate (18) cannot be improved since it is attained on an arbitrary fourth-degree polynomial. It is easy to prove this by analogy with what we have done for estimate (9).

To estimate the computing error of the result, obtained from formula (16), we assume that the values of the function are specified with an accuracy of ε_1 and the values of the derivatives with an accuracy of ε_2. Then the computing error is

$$\Delta_2 \leqslant (b-a)\,\varepsilon_1 + \frac{(b-a)^2}{6}\varepsilon_2. \tag{19}$$

Example 5. Use the quadrature formula (16) to calculate the integral $\int\limits_0^1 \frac{dx}{1+x}$. Evaluate the error of the approximate value obtained.

△ Calculating the required values of the integrand function and its derivatives with the use of formula (16), we find that

$$I_4 = 1 \cdot \left(\frac{1+0.5}{2} + 1 \cdot \frac{-1+0.25}{12}\right) = 0.6875.$$

We use formula (18) to evaluate the error of the method, for which purpose we first find the maximum of the absolute value of the fourth derivative of the integrand function $M_4 = 24$:

$$\Delta_1 \leqslant \frac{1^5}{720}\cdot 24 \cong 0.034.$$

The computing error is evidently zero since the values of the function and those of the derivatives have been calculated with the absolute accuracy.

Thus, rounding off the approximate values of the integral and the error, we finally obtain $\int\limits_0^1 \frac{dx}{1+x} = 0.69 \pm 0.04$. ▲

Up till now, in all the quadrature formulas we have considered, the quadrature nodes were fixed. We shall now consider the case when the position of all the nodes as well as all the weight factors are assumed to be free parameters. For the computations not to be very complicated but at the same time not trivial, we shall seek the value of the integral in the form

$$\int\limits_a^b f(x)\,dx = (b-a)\,[A_1 f(x_1) + A_2 f(x_2)]. \tag{20}$$

To determine the four free parameters A_1, A_2, x_1 and x_2, we require that formula (20) should be absolutely accurate for all polynomials of degree zero, one, two and three. By virtue of linearity of the operation of integration and the right-hand side of relation (20), for the quadrature formula (20) to be exact for all third-degree polynomials, it is necessary and sufficient that it be exact for the functions 1, x, x^2 and x^3. Consequently, there must hold relations

$$\left\{\begin{aligned} &A_1 + A_2 = 1,\\ &x_1A_1 + x_2A_2 = \frac{1}{2}(a+b),\\ &x_1^2A_1 + x_2^2A_2 = \frac{1}{3}(a^2+ab+b^2),\\ &x_1^3A_1 + x_2^3A_2 = \frac{1}{4}(a^3+a^2b+ab^2+b^3),\end{aligned}\right.$$

which constitute a nonlinear system of equations with respect to the parameters x_1, x_2, A_1 and A_2 being determined.

The solution of this system is

$$A_1 = A_2 = \frac{1}{2},$$

$$x_1 = \frac{b+a}{2} - \frac{b-a}{2} \cdot \frac{1}{\sqrt{3}}, \quad x_2 = \frac{b+a}{2} + \frac{b-a}{2} \cdot \frac{1}{\sqrt{3}}. \tag{21}$$

Thus the quadrature formula (20) assumes the form

$$I \equiv \int_a^b f(x)\,\mathrm{d}x \cong \frac{b-a}{2}\left[f\left(\frac{b+a}{2} - \frac{b-a}{2}\frac{1}{\sqrt{3}}\right) + f\left(\frac{b+a}{2} + \frac{b-a}{2}\frac{1}{\sqrt{3}}\right)\right] \equiv I_2. \tag{22}$$

Formulas of the kind when not only weight factors but also nodal points are not fixed in advance are known as *Gaussian formulas.*

We seek an expression for the remainder of formula (22). For this purpose we represent the function being integrated as the sum of a third-degree Hermite interpolating polynomial with two double nodes x_1 and x_2, defined by relations (21), and the remainder. Integrating the right-hand and left-hand sides of this representation on the interval $[a, b]$, we obtain

$$\int_a^b f(x)\,\mathrm{d}x = \int_a^b H_3(x)\,\mathrm{d}x + \int_a^b \frac{(x-x_1)^2(x-x_2)^2}{4!} f(\bar{\eta})\,\mathrm{d}x,$$

$$\bar{\eta} \in (a, b).$$

The first term on the right-hand side yields the right-hand side of the quadrature formula (22) since this formula is exact for all third-degree polynomials and, consequently, for $H_3(x)$ as well. The second term on the right-hand side yields the remainder of formula (22). Using the second theorem of the mean and carrying out the integration, we have

$$R_2[f] = I - I_2 = \frac{(b-a)^5}{4320} f^{\mathrm{IV}}(\eta), \quad \eta \in (a, b). \tag{23}$$

Consequently, the estimate of the error is expressed by the relation

$$\Delta_1 = |I - I_2| \leqslant \frac{(b-a)^5}{4320} M_4, \qquad (24)$$

where $M_4 = \max\limits_{[a,\, b]} |f^{\text{IV}}(x)|$.

The estimate obtained cannot be improved since it is attained on an arbitrary fourth-degree polynomial. We can easily show this by means of direct computations as we did for estimate (9).

If the values of the nodes in formula (22) are practically exact and the values of the function have been found with an absolute error ε, then we shall get the same expression for the computing error with the use of formula (2) as for the computing error with the use of formulas (4) and (11):

$$\Delta_2 \leqslant (b-a)\,\varepsilon. \qquad (25)$$

Example 6. Use the quadrature formula (22) to calculate the integral $\int\limits_0^1 \frac{dx}{1+x}$. Evaluate the error of the approximate value obtained.

△ First of all we find the nodes of the quadrature formula:

$$x_1 = \frac{1}{2}\left(1 - \frac{1}{\sqrt{3}}\right) = 1.2113249\ \ldots,$$

$$x_2 = \frac{1}{2}\left(1 + \frac{1}{\sqrt{3}}\right) = 0.7886751\ .$$

Having calculated the required values of the integrand function with an accuracy to within three valid digits in the narrow sense, we use formula (22):

$$\int\limits_0^1 \frac{dx}{1+x} \cong \frac{1}{2}(0.826 + 0.559) = 0.6925.$$

We find the error of the method from formula (24), for which purpose we use the value of the maximum of the absolute value of the derivative $M_4 = 24$ found in Example 3:

$$\Delta_1 \leqslant \frac{1^5}{4320} \cdot 24 \approx 0.0056,$$

We can find the computing error from formula (24) taking into account that the accuracy of the calculation of the values of the function being integrated is equal to 0.0005.

Thus the total error is $\Delta = \Delta_1 + \Delta_2 = 0.0061$.

Finally, rounding off the approximate value of the integral, we have $\int\limits_0^1 \frac{dx}{1+x} = 0.692 \pm 0.007$. ▲

The main purpose of this section is to show by simple examples how to derive various formulas of numerical integration. We have not naturally considered all the methods of constructing formulas. All the same, the examples given are typical so that using them, the student can construct a specific quadrature formula which suits best of all the practical problem posed.

8.5. Newton-Cotes Quadrature Formulas

In this section we discuss formulas of numerical integration which are more complicated in structure. Up till now we considered interpolation methods of numerical integration including, in a definite sense, the rectangular formula. This means that the integrand function was approximated by an interpolating polynomial. If the function being integrated is smooth enough and the interval of integration is finite, we can get sufficiently good results. On the other hand, it is hardly possible to attain a close approximation of a function by a polynomial if the function itself or its derivatives of low orders have peculiarities. In such cases it is expedient to represent the integrand function as the product of two factors $\rho(x) f(x)$, which must possess the following three properties. First, the weight factor $\rho(x)$ must reflect all the peculiarities of the function being integrated and, second, the moments

$$m_k = \int\limits_a^b \rho(x) x^k \, dx \quad (k = 0, 1, \ldots), \tag{1}$$

where $[a, b]$ is an integration interval, must be calculated by analytical methods. Third, the error of approximation of the function $f(x)$ by a polynomial must be small.

Let us now construct the quadrature formulas themselves. We shall construct them in the same form as before:

$$I \equiv \int_a^b \rho(x) f(x)\,dx \cong (b-a) \sum_{i=1}^{n} A_i f(x_i) \equiv I_n. \qquad (2)$$

In the general case, as we mentioned before, formula (2) has $2n$ free parameters which are the quadrature nodes x_i and the weight factors A_i. We assume the number n to be fixed. The choice of free parameters is defined by the same requirements that are imposed upon a quadrature formula by the conditions of a practical problem. These requirements may be, for instance, the maximum possible accuracy, the minimum computing error, the fixing of some (and, maybe, all) weight factors or quadrature nodes.

We begin with a relatively simple case when the nodes are specified in advance and we can vary only the choice of the weight factors A_i. The idea of interpolating quadratures is that we approximate the function f by an interpolating polynomial in Lagrange's form of degree $n-1$ using n different nodes x_i:

$$f(x) = L_{n-1}(x) + R_{n-1}(x),$$

integrate the right-hand and left-hand sides of this relation on the interval $[a, b]$ first multiplying them by the weight function $\rho(x)$:

$$\int_a^b \rho(x) f(x)\,dx = \int_a^b \rho(x) L_{n-1}(x)\,dx + \int_a^b \rho(x) R_{n-1}(x)\,dx, \quad (3)$$

and transform the first term on the right-hand side of this relation, for which purpose we replace L_{n-1} by its explicit expression and interchange the operations of integration and summation:

$$\int_a^b \rho(x) f(x)\,dx = \sum_{i=1}^{n} \left(\int_a^b \rho(x) \frac{\omega_{n-1}(x)}{(x-x_i)\,\omega'_{n-1}(x_i)}\,dx \right) f_i. \qquad (4)$$

Here $\omega_{n-1}(x) = \prod_{h=1}^{n} (x-x_h)$. The first factor under the sign of the sum is numerical coefficient which is proportional to the length of the integration interval and depends only on the position of the nodes and the properties of the function $\rho(x)$ (but not on the function f).

Assuming now that the second term on the right-hand side of relation (3) is small, we get the approximate quadrature formula (2) with the specified nodes x_i and coefficients A_i defined as follows:

$$A_i = \frac{1}{b-a}\int_a^b \rho(x)\,\frac{\omega_{n-1}(x)}{(x-x_i)\,\omega'_{n-1}(x_i)}\,dx \quad (i=1,\ 2,\ \ldots,\ n). \qquad (5)$$

The quadrature formula (2) thus constructed is known as an *interpolation formula.*

Let us now estimate the error of formula (2) with coefficients (5). To do this, we integrate the remainder of the interpolation formula $R_{n-1}(x) = \frac{\omega_{n-1}(x)}{n!}f^{(n)}(\bar{\eta})$. Substituting this expression for $R_{n-1}(x)$ into the second term of relation (3), we obtain

$$R_{n-1}[f] = I - I_n = \frac{1}{n!}\int_a^b \rho(x)\,\omega_{n-1}(x)\,f^{(n)}(\bar{\eta})\,dx,\quad \bar{\eta}\in(a,\ b).$$

If the function f has a continuous derivative of order n on the interval of integration and the product $\rho(x)\,\omega_n(x)$ retains sign on this interval, then we can get the following expression for the remainder:

$$R_{n-1}[f] = \frac{f^{(n)}(\eta)}{n!}\int_a^b \rho(x)\,\omega_{n-1}(x)\,dx,\quad \eta\in(a,\ b), \qquad (6)$$

and, consequently, the estimate of the error of the quadrature assumes the form

$$\Delta_1 \leqslant \frac{M_n}{n!}\left|\int_a^b \rho(x)\,\omega_{n-1}(x)\,dx\right|, \qquad (7)$$

where $M_n = \max\limits_{[a,\,b]} |f^{(n)}(x)|$. Under the conditions indicated above this estimate is the best.

Now if the product $\rho(x)\,\omega_{n-1}(x)$ does not retain sign on the integration interval, then we get only a rough estimate of the error $\Delta_1 \leqslant \frac{M_n}{n!} \int\limits_a^b |\rho(x)\,\omega_{n-1}(x)|\,dx$,

which may turn out to be far from the optimal one. Therefore, in such cases other reasons are taken into consideration when an explicit expression is constructed for the remainder and the error. We shall discuss one of these techniques later on when we study Simpson's rule.

Example 1. Construct the quadrature formula (2) for the interval $[-1, 1]$ with nodes $x_1 = -1$, $x_2 = 0$, $x_3 = 1$ and a weight function $\rho(x) = (1 - x^2)^{-1/2}$.

△ In essence, we have to determine the coefficients A_i ($i = 1, 2, 3$) appearing in formula (2). Using expression (5) for the required coefficients, we have

$$2A_1 = \int\limits_{-1}^{1} \frac{(x-0)(x-1)}{(-1-0)(-1-1)} \cdot \frac{dx}{\sqrt{1-x^2}} = \frac{\pi}{4},$$

$$2A_2 = \int\limits_{-1}^{1} \frac{(x+1)(x-1)}{(0+1)(0-1)} \cdot \frac{dx}{\sqrt{1-x^2}} = \frac{\pi}{2},$$

$$2A_3 = \int\limits_{-1}^{1} \frac{(x+1)(x-0)}{(1+1)(1-0)} \cdot \frac{dx}{\sqrt{1-x^2}} = \frac{\pi}{4}.$$

Thus the required formula has the form

$$\int\limits_{-1}^{1} f(x) \frac{dx}{\sqrt{1-x^2}} \approx \frac{\pi}{4}[f(-1)+2f(0)+f(1)]. \ \blacktriangle$$

Example 2. Construct the quadrature formula (2) for the interval $[0, 1]$ with nodes $x_1 = 0$, $x_2 = 0.5$, $x_3 = 1$ and weight function $\rho(x) = \ln x$.

△ As in the preceding example, we use formula (5) to find the coefficients:

$$A_1 = \int\limits_0^1 \ln x \frac{(x-0.5)(x-1)}{(0-0.5)(0-1)}\,dx = -\frac{17}{36},$$

$$A_2 = \int\limits_0^1 \ln x \frac{(x-0)(x-1)}{(0.5-0)(0.5-1)}\,dx = -\frac{20}{36},$$

$$A_3 = \int\limits_0^1 \ln x \frac{(x-0)(x-0.5)}{(1-0)(1-0.5)}\,dx = \frac{1}{36}.$$

Thus

$$\int_0^1 \ln x \, f(x)\,dx \cong -\frac{1}{36}[17f(0)+20f(0.5)-f(1)]. \ \blacktriangle$$

Note the characteristic features of the examples considered. In Example 1 the symmetry of the nodes and the evenness of the weight function with respect to the midpoint of the interval led to the symmetry of the coefficients of the quadrature formula. In Example 2, despite the symmetry of the nodes, the symmetry of the coefficients is violated, which is the consequence of the absence of symmetry (evenness) of the weight function.

In practical computations of especial interest is the case when the nodes of the quadrature formula are given as equispaced points of the interval $[a, b]$: $x_i = a + (i-1)h$ $(i = 1, 2, \ldots, n)$ and the weight function $\rho(x)$ is identically equal to unity. On these assumptions we can transform formula (2) as follows:

$$\int_a^b f(x)\,dx \cong (b-a)\sum_{i=1}^{n} H_i f_i. \tag{8}$$

For different n we get different *quadrature formulas of Newton-Cotes*. The coefficients H_i known as *Cotes' coefficients* can be found from relation (5):

$$H_i = A_i = \frac{(-1)^{n-i}}{(n-1)(i-1)!\,(n-i)!}\int_1^n \frac{(t-1)\ldots(t-n)}{t-i}\,dt,$$

$$n>1,\ i=1,\ 2,\ \ldots,\ n;\quad 0!=1. \tag{9}$$

These coefficients possess the following properties useful for calculations.

1°. *The symmetrical coefficients (the first and the nth, the second and the $(n-1)$ th, . . .) are equal to one another*:

$$H_i = H_{n+1-i}.$$

☐ We replace i in expression (9) by $n+1-i$:

$$H_{n+1-i} = \frac{(-1)^{i-1}}{(n-1)(n-i)!\,(i-1)!}\int^n \frac{(t-1)\ldots(t-n)}{t-n-1+i}\,dt.$$

Passing to a new variable $q = n + 1 - t$ under the integral sign and carrying out simple transformations, we obtain

$$H_{n+1-i} = \frac{(-1)^{n-i}}{(n-1)(n-i)!\,(i-1)!} \int_1^n \frac{(q-n)\ldots(q-1)}{q-i}\,dq,$$

and this coincides with expression (9) for the coefficients H_i. ■

2°. *The sum of all the coefficients is equal to unity:*

$$\sum_{i=1}^{n} H_i = 1.$$

□ The validity of this property follows immediately from formula (8) if we set $f(x) = 1$ since the remainder of this formula, defined by expression (6), is zero for $f(x) = 1$. ■

Table 8.1 gives the values of Cotes' coefficients for $n = 2, 3, 4, 5, 6$.

Table 8.1

n \ i	1	2	3	4	5	6
2	1/2	1/2				
3	1/6	4/6	1/6			
4	1/8	3/8	3/8	1/8		
5	7/90	32/90	12/90	32/90	7/90	
6	19/288	75/288	50/288	50/288	75/288	19/288

We shall consider in more detail a significant special case of the quadrature formula (8) resulting at $n = 3$. To construct this formula, we could use the data from Table 8.1 but we shall perform the requisite calculations by way of an example. Employing formula (9), we obtain

$$H_1 = \frac{(-1)^2}{2\cdot 1\cdot 2}\cdot \int_1^3 (t-2)(t-3)\,dt = \frac{1}{6}.$$

Next, using relation (10), we find that

$$H_3 = H_1 = 1/6, \quad H_2 = 1 - (H_1 + H_3) = 2/3.$$

Thus the required quadrature formula, known as *Simpson's formula*, has the form

$$I \equiv \int_a^b f(x)\,\mathrm{d}x \cong \frac{b-a}{6}\left[f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right] \equiv I_3. \tag{11}$$

By virtue of its construction (approximation of the integrand by a second-degree polynomial), this formula is exact for all polynomials of degree zero, one and two. We could try to obtain an expression for the remainder directly from relation (6). However, Simpson's formula possesses the so-called **property of increased accuracy** meaning that it is exact not only for second-degree polynomials but also for third-degree polynomials. Since the operation of integration is linear, it is sufficient to establish an exact equality of the right-hand and left-hand sides of formula (11) for the simplest third-degree polynomial x^3 in order to prove this statement. Indeed, calculating the left-hand side of formula (11) for $f = x^3$, we have

$$\int_a^b x^3\,\mathrm{d}x = \frac{b^4-a^4}{4}.$$

On the other hand, calculating the right-hand side of formula (11), we obtain

$$\frac{b-a}{6}\left[a^3+4\left(\frac{b+a}{2}\right)^3+b^3\right] = \frac{b^4-a^4}{4},$$

and this is what we wished to prove.

We shall take advantage of the statement we have proved to construct the remainder of Simpson's formula. We represent the function f as the sum of Hermite's interpolating polynomial with simple nodes a and b, a double node $(a+b)/2$ and a remainder

$$f(x) = H_3(x) + (x-a)\left(x - \frac{a+b}{2}\right)^2 (x-b)\frac{f^{\mathrm{IV}}(\bar{\eta})}{4!},$$

$$\bar{\eta} \in (a, b).$$

We integrate the right-hand and left-hand sides of this relation on the interval $[a, b]$. By virtue of what we have just proved, the integral of Hermite's polynomial yields the right-hand side of formula (11) and the integral of the second term yields its remainder

$$R_3[f] = I - I_3 = -\frac{(b-a)^5}{2880} f^{\mathrm{IV}}(\eta), \quad \eta \in (a, b). \tag{12}$$

Consequently, we can represent the estimate of the error as

$$\Delta_1 \leqslant |I - I_3| \leqslant \frac{(b-a)^5}{2880} \cdot M_4, \tag{13}$$

where $M_4 = \max\limits_{[a,b]} |f^{\mathrm{IV}}(x)|$. This estimate cannot be improved since it is attained, say, for the function $f = x^4$.

Example 3. Use Simpson's rule to calculate the integral $\int\limits_0^1 \frac{dx}{1+x}$. Estimate the error of the approximate value obtained.

△ Calculating the needed values of the integrand at the points $x_1 = 0$, $x_2 = 0.5$, $x_3 = 1$, we substitute them into formula (11):

$$I_3 = \frac{1}{6}(1 + 4 \cdot 0.667 + 0.5) = 0.6947.$$

Taking into account that $M_4 = \max\limits_{[0,1]} \left| \left(\frac{1}{1+x}\right)^{\mathrm{IV}} \right| = 24$ and using formula (13), we find that the error of the method $\Delta_1 \leqslant 0.0084$.

Let us find the computing error:

$$\Delta_2 \leqslant \frac{1}{6} \cdot (0 + 4 \cdot 0.0005 + 0) = 0.00034.$$

Adding the errors together and rounding off the result, we obtain $\int\limits_0^1 \frac{dx}{1+x} = 0.695 \pm 0.01$. ▲

The property of increased accuracy described above is inherent in all quadrature formulas of type (8) constructed with the use of an odd number of nodes. This property is a direct consequence of the "symmetry" of the quadrature, i.e. the equality of the symmetric coefficients $H_i = H_{n+1-i}$, and the linearity of the operation of computing the value of a function and that of integration. To obtain a precise estimate of the remainder of such a formula, it is necessary to approximate the integrand by Hermite's interpolating polynomial with a double central node.

Returning to the quadrature formula (2), we can state that when the coefficients are symmetrical ($A_i = A_{n+1-i}$) and the weight function is even, this quadrature constructed with the use of $2k + 1$

nodes is exact for all polynomials of degree $2k+1$ since it is exact for any function f which is odd with respect to the midpoint of the interval of integration. Indeed, on the one hand, $\int\limits_a^b \rho(x) f(x)\,dx = 0$ for functions of this kind, and, on the other hand, $\sum\limits_{i=1}^{2k+1} A_i f_i = 0$ by virtue of the symmetry of A_i and the oddness of f.

Example 4. Using the quadrature formula constructed in Example 1, calculate the integral $\int\limits_{-1}^{1} \frac{\cos x\,dx}{\sqrt{1-x^2}}$ and estimate the error.

△ We calculate the needed values of the function $f = \cos x$: $\cos(-1) = 0.540$, $\cos 0 = 1$, $\cos 1 = 0.540$. Substituting these values into the quadrature formula, we have

$$\int\limits_{-1}^{1} \frac{\cos x\,dx}{\sqrt{1-x^2}} \cong \frac{\pi}{4}(0.540+2\cdot 1+0.540)=2.419.$$

We seek now an expression for the remainder of the quadrature formula given in Example 1. Since the weight function is even and the nodes are symmetric and odd in number, this formula possesses the property of increased accuracy, i.e. it is exact for all third-degree polynomials. Therefore, by analogy with what we did for Simpson's formula, we use Hermite's interpolating polynomial with a double central node $x_2 = 0$ and get the following expression for the remainder:

$$R_3[f] = \frac{f^{\mathrm{IV}}(\eta)}{4!} \int\limits_{-1}^{1} \frac{(x+1)x^2(x-1)}{\sqrt{1-x^2}}\,dx = -\frac{\pi}{192} f^{\mathrm{IV}}(\eta); \ \eta \in (-1, 1).$$

Consequently, the error of the method is

$$\Delta_1 \leqslant \frac{\pi}{192} \max_{[-1,\,1]} |\cos x| = 0.017.$$

Next, since we have calculated the first and the third value of the cosine with an error of 0.0005 and the second with an absolute accuracy, we get the following expression for the computing error:

$$\Delta_2 \leqslant \frac{\pi}{4}(0.0005+2\cdot 0+0.0005)=0.0008.$$

Summing up the errors and carrying out the necessary roundings-off, we finally obtain

$$\int\limits_{-1}^{1} \frac{\cos x\,dx}{\sqrt{1-x^2}} = 2.42 \pm 0.02. \ ▲$$

In conclusion we shall discuss one more technique of determining the coefficients A_i appearing in the quadrature formula (2) with specified nodes. We require that formula (2) should be as accurate for the functions 1, x, x^2, . . . as possible. In this case, we get the following system of equations for the coefficients A_i:

$$m_k = (b-a)\sum_{i=1}^{n} x_i^k A_i \quad (k=0,\ 1,\ \ldots,\ N). \qquad (14)$$

If the nodes x_i do not coincide and $N = n - 1$, then the determinant of system (14) is a Vandermonde determinant and the solution of this system (the collection of coefficients A_i) exists and is unique.

We shall prove that such a method of determining quadrature coefficients is equivalent to that described above, i.e. that formulas (5) yield the same values for A_i as system (14). To prove this fact, we substitute expressions (5) for the coefficients A_i into the right-hand sides of equations (14) and interchange the operations of summation and integration:

$$m_k = \int_a^b \rho(x)\left[\sum_{i=1}^{n} \frac{\omega_{n-1}(x)}{(x-x_i)\,\omega'_{n-1}(x_i)}\, x_i^k\right] dx.$$

The expression in brackets is Lagrange's interpolating polynomial of a degree not higher than $n - 1$ for the function x^k $(k < n)$. The remainder of such a polynomial is evidently zero $\left(\frac{d^n}{dx^n}(x^k) = 0 \text{ for } k < n\right)$, and therefore

$$m_k = \int_a^b \rho(x)\, x^k\, dx \quad (k=0,\ 1,\ \ldots,\ n-1).$$

It follows from the identities obtained and the uniqueness of the solution of system (14) that both methods of constructing the quadrature formulas (2) are equivalent.

By way of an example, we shall construct a quadrature formula of the form

$$\int_a^b f(x)\,dx \cong (b-a)\left[A_1 f'(a) + A_2 f\left(\frac{a+b}{2}\right) + A_3 f'(b)\right].$$

To construct such a quadrature formula, we require that it should be exact for all polynomials of degree zero, one and two. For this purpose, by virtue of the linearity of the operations of integration, calculation of the values of the function and differentiation, it is sufficient to require that the quadrature formula should be accurate for 1, x, x^2. Thus let $f(x) = 1$ and then

$$(b-a) = (b-a)(A_1\cdot 0 + A_2\cdot 1 + A_3\cdot 0), \text{ i.e. } A_2 = 1.$$

Assume now that $f(x) = x$ and then

$$\frac{b+a}{2} = A_1\cdot 1 + 1\cdot\frac{b+a}{2} + A_3\cdot 1, \text{ i.e. } A_1 = -A_3.$$

Finally, setting $f(x) = x^2$, we get an equation

$$\frac{a^2+ab+b^2}{3} = A_1\cdot 2a + 1\cdot\left(\frac{a+b}{2}\right)^2 - A_1\cdot 2b,$$

solving which we find that $A_1 = -\frac{b-a}{24}$.

Thus the required formula has the form

$$I \equiv \int_a^b f(x)\,dx$$

$$\approx \frac{b-a}{24}\Big[-(b-a)f'(a) + 24f\left(\frac{a+b}{2}\right)$$

$$+ (b-a)f(b)\Big] \equiv I_3.$$

Note that this formula also possesses the property of increased accuracy as, for instance, Simpson's formula. Here we give, without deriving, the expression for the remainder of the formula obtained

$$R_3[f] = I - I_3 = -\frac{7}{5760}(b-a)^5 f^{\mathrm{IV}}(\eta),\ \eta\in(a, b).$$

We invite the reader to prove the last relation independently as an exercise.

8 6. Quadrature Formulas of the Highest Algebraic Degree of Accuracy

We again consider the quadrature formula

$$I \equiv \int_a^b \rho(x) f(x)\, dx \cong (b-a) \sum_{i=1}^{n} A_i f(x_i) \equiv I_n. \quad (1)$$

As we have noted, in the general case not only the coefficients A_i but also the nodes x_i are arbitrary parameters that must be determined in accordance with the requirements which relation (1) must satisfy. Since the total number of free parameters is $2n$, we can expect that imposing $2n$ conditions on relation (1), we shall get a system of equations for determining these parameters. We can certainly not find out for sure, in the general case, whether the system has a solution at all, and if it has, then whether it is unique. We can answer these questions only after considering the concrete requirements the quadrature formula (1) must satisfy. It is evidently expedient to connect these requirements with the value of the error of formula (1) and to try to choose the nodes and the coefficients so as to minimize, in a certain sense, the absolute value of the remainder. We pointed out in the preceding section that the estimates of the errors of the quadrature formulas were attained on polynomials whose degree exceeded by unity the maximum degree of the polynomial for which the corresponding quadrature was exact. It is natural therefore to try and increase the degree of the polynomial for which formula (2) would be absolutely exact. Such a posing of the problem generates the following optimization problem.

We have to construct a quadrature formula of type (1) with a fixed n, accurate for an arbitrary polynomial of degree r as high as possible.

We shall show that this problem is equivalent to the problem of constructing the quadrature formula (1) which is accurate for all functions x^k $(k = 0, 1, \ldots, r)$. Indeed, assume that we have an arbitrary polynomial

$P_r(x) = \sum_{k=0}^{r} a_k x^k$. Then

$$\begin{aligned} R_n[P_r] &= I[P_r] - I_n[P_r] \\ &= \sum_{k=0}^{r} a_k \left(\int_a^b \rho(x) x^k \, dx - \sum_{i=1}^{n} A_i x_i^k \right) \\ &= \sum_{k=0}^{r} a_k R_n[x^k]. \end{aligned} \tag{2}$$

From this, by virtue of the arbitrariness of a_k (the polynomial P_r is arbitrary), we find that for $R_n[P_r] = 0$ it is necessary that $R_n[x^k] = 0$ $(k = 0, 1, \ldots, r)$. The sufficiency is a direct consequence of the linearity of the remainder of formula (1) with respect to the function f and is obvious by virtue of the same relation (2).

We thus arrive at a system of $r + 1$ equations for $2n$ unknown parameters A_i and x_i:

$$m_k = \int_a^b \rho(x) x^k \, dx = (b-a) \sum_{i=1}^{n} A_i x_i^k \quad (k = 0, 1, \ldots, r). \tag{3}$$

It is natural to try and solve this system, i.e. construct a quadrature formula for $r + 1 = 2n$ (the number of equations is equal to the number of unknowns). The theorems presented below substantiate the expediency of this attempt.

We shall first formulate without proof an auxiliary theorem which characterizes the properties of orthogonal polynomials.

Theorem 1. *Assume that*: (1°) $\rho(x) > 0$ *almost everywhere on* $[a, b]$, (2°) $P_{n-1}(x)$ *is an arbitrary polynomial of degree not higher than* $n - 1$. *Then there is a polynomial*

$$\Psi_n(x) = (x - x_1)(x - x_2) \ldots (x - x_n) \tag{4}$$

orthogonal to $P_{n-1}(x)$ *with weight* $\rho(x)$, *i.e. such that*

$$\int^b \rho(x) \Psi(x) P_{n-1}(x) \, dx = 0, \tag{5}$$

all its roots $x_1, x_2, \ldots, x_n$ being distinct and lying within the interval $[a, b]$, and this polynomial is unique.

We shall now consider theorems which make it possible to find directly the nodes and coefficients of formula (1) which is accurate for all polynomials of degree $r = 2n - 1$.

Theorem 2. *For formula* (1) *to be accurate for polynomials of degree $2n - 1$, it is necessary and sufficient that:* (1°) *the nodes x_i be roots of the polynomial Ψ_n defined by relation* (5), (2°) *the weight factors A_i be defined by relation* (5) *from* 8.5.

□ We begin with the necessity of condition 2°. If formula (1) is accurate for all polynomials of degree $2n - 1$, then it is also accurate for polynomials of any lower degree, the degree $n - 1$ inclusive; then this formula is interpolating and we get a unique collection of coefficients A_i defined by formula (5) from 8.5.

Let us consider a polynomial $Q_{2n-1} = \omega_{n-1}(x) P_{n-1}(x)$. Using the fact that formula (1) is accurate for the polynomial $Q_{2n-1}(x)$, we obtain condition 1°:

$$\int_a^b \rho(x)\,\omega_{n-1}(x)\,P_{n-1}(x)\,dx = (b-a)\sum_{i=1}^{n} A_i Q_{2n-1}(x_i) = 0.$$

The last relation follows from the fact that $\omega_{n-1}(x_i) = 0$ for all $i = 1, 2, \ldots, n$.

We shall verify the sufficiency of conditions 1 and 2°. We represent the arbitrary polynomial $Q_{2n-1}(x)$ of degree $2n - 1$ as

$$Q_{2n-1} = \Psi_n P_{n-1}(x) + S_{n-1}(x),$$

where P_{n-1} and S_{n-1} are the quotient and the remainder of the division of the polynomial Q_{2n-1} by the polynomial Ψ_n respectively, P_{n-1} being a polynomial of degree $n - 1$ and S_{n-1}, a polynomial of degree not higher than $n - 1$. Next we have

$$\int_a^b \rho(x)\,Q_{2n-1}\,dx = \int_a^b \rho(x)\,\Psi_n P_{n-1}\,dx + \int_a^b \rho(x)\,S_{n-1}\,dx.$$

The first term on the right-hand side is zero by virtue of the orthogonality of Ψ_n and P_{n-1} (condition 1°) and for the second term formula (1) is accurate by virtue of condition 2°. Therefore, taking into account that $Q_{2n-1}(x_i) = S_{n-1}(x_i)$ since $\Psi_n(x_i) = 0$, we finally obtain

$$\int_a^b \rho(x) Q_{2n-1}\,dx = (b-a) \sum_{i=1}^{n} A_i Q_{2n-1}(x_i). \ \blacksquare$$

The next question is whether it is possible to construct the quadrature formula (1) accurate for all polynomials of degree higher than $2n-1$. The following theorem answers this question.

Theorem 3. *Let $\rho(x) > 0$ almost everywhere on $[a, b]$. Then there is no quadrature formula of type* (1) *accurate for all polynomials of degree $2n$.*

□ We consider a polynomial $Q_{2n}(x) = (x-x_1)^2 \ldots (x-x_n)^2 = \omega_{n-1}^2(x)$. Then the left-hand side of relation (1) is $\int_a^b \rho(x)\,\omega_{n-1}^2(x)\,dx > 0$ whereas its right-hand side is $\sum_{i=1}^{n} A_i \omega_{n-1}^2(x_i) = 0$, and this proves the theorem. ■

Thus the quadrature formula (1) with nodes x_i defined by relation (5) and coefficients A_i defined by relation (5) from 8.5 is accurate for any polynomial of degree not higher than $2n-1$. It is known as a *quadrature formula of the highest algebraic degree of accuracy* (or the *Gaussian quadrature formula*).

The Gaussian formula possesses a property useful from the point of view of computing error: all coefficients A_i $(i = 1, 2, \ldots, n)$ are positive for any n.

To prove this statement, we consider a function $\left[\frac{\Psi_n(x)}{x-x_p}\right]^2$ which is a polynomial of degree $2n-2$ and which vanishes at all nodes $x_i \neq x_p$. The Gaussian formula is accurate for this function and therefore

$$\int_a^b \rho(x) \frac{\Psi_n^2(x)}{(x-x_p)^2}\,dx = (b-a) A_p [\Psi'(x_p)]^2.$$

It follows immediately that $A_p > 0$. Since ρ is arbitrary, all quadrature coefficients A_i evidently exceed zero.

We shall now estimate the error of the Gaussian quadrature formula.

Theorem 4. *Assume that:* (1°) $\rho(x) > 0$ *almost everywhere on* $[a, b]$, (2°) $f(x) \in C^{2n}[a, b]$, (3°) x_i and A_i $(i = 1, 2, \ldots, n)$ *are the nodes and the coefficients of the Gaussian quadrature formula respectively. Then there is a point* $\xi \in (a, b)$ *such that*

$$R_n[f] = I - I_n = \frac{f^{(2n)}(\xi)}{(2n)!} \int_a^b \rho(x)\, \omega_{n-1}^2(x)\, dx. \qquad (6)$$

☐ We represent the function being integrated as

$$f(x) = H_{2n-1}(x) + \frac{f^{(2n)}(\bar{\eta})}{(2n)!} \omega_{n-1}^2(x), \quad \bar{\eta} \in (a, b),$$

where $H_{2n-1}(x)$ is the Hermitian interpolating polynomial of degree not higher than $2n - 1$ with double nodes x_i $(i = 1, 2, \ldots, n)$. Evidently, the Gaussian formula is exact for $H_{2n-1}(x)$. Therefore

$$I = I_n + \frac{1}{(2n)!} \int_a^b \rho(x) f^{(2n)}(\bar{\eta})\, \omega_{n-1}^2(x)\, dx.$$

Applying the second theorem of the mean to the second term on the right-hand side and transferring I_n to the left-hand side, we get the required relation (6). ■

In practice, the Gaussian quadrature is usually obtained as follows. We first find the roots of the polynomial $\Psi_n(x)$ which is orthogonal to all polynomials of degree lower than n. Then we construct a system of linear equations for the coefficients A_i assuming that the quadrature formula is valid for the functions $1, x, \ldots, x^{n-1}$:

$$m_k \equiv \int_a^b \rho(x)\, x^k\, dx = \sum_{i=1}^{n} A_i x_i^k.$$

Solving this system, we find A_i.

Formula (22) constructed in 8.4 is the simplest illustration of the Gaussian quadrature.

8.7. Compounded Quadrature Formulas

Formulas with a large number of equispaced nodes are not widely used. There are various reasons for this. First, for functions which have a singularity not on the integration interval itself but in its vicinity, the remainder of a quadrature formula of a high order is large, as a rule. And what is more, when the nodal points are equally spaced, even for some analytic functions, it is possible that $R_n[f] = I - I_n \to \infty$ as $n \to \infty$. Second, for $n \geqslant 10$ there are negative coefficients among H_i, and this leads to an essential increase in the expected computing error. Indeed, assume that all the values f_i of the function x appearing in formula (8) from 8.5 are given with the same accuracy ε. Then we can estimate the total computing error by the quantity $\Delta_2 = (b - a)\,\varepsilon \sum\limits_{i=1}^{n} |H_i|$. Since the sum of all Cotes coefficients is equal to unity, it follows that if there are negative coefficients among them, the value of Δ_2 is increased. Table 8.2 demonstrates the rate of growth of Δ_2.

Table 8.2

n	10	15	20
$\sum\limits_{i=1}^{n} \lvert H_i \rvert$	≈ 3	≈ 8	≈ 560

Formulas of the Gaussian type possess definite advantages as compared with the Newton-Cotes formulas since they are free from the drawbacks described above. However, being formulas of the highest algebraic degree of accuracy, the Gaussian formulas have, for large n, a remainder which is proportional to the derivative of the function being integrated whose order is high. This is an essential drawback as concerns the integration of functions which do not possess continuous derivatives of higher orders or of empirically constructed functions. Besides this, even for convergent quadrature processes constructed by means of Gaussian formulas it is usually not known in advance how high n must be for the convergence to begin in practice, i.e. it is not known in advance how high n must be for the required accuracy to be ensured.

For these reasons, the so-called *compounded* (*generalized*) *formulas* are usually preferred in practical calculations. The essence of these formulas is that the interval of integration $[a, b]$ is divided into several subintervals, some quadrature formula with small n is applied to each subinterval and the results are added together.

The expediency of such an approach is based on the following arguments. For many quadrature formulas, including those considered in this chapter, the remainder is proportional to the power of the length of the integration interval. Assume, for instance, that the application of some chosen quadrature formula to the interval $[a, b]$ yields the following expression for the remainder:

$$R[f] = (b-a)^k \varphi(a, b), \tag{1}$$

where $\varphi(a, b)$ is a slowly varying function of the integration interval.

Dividing the original interval into m equal subintervals and applying the same quadrature formula to each of them, we find that on each subinterval the remainder is approximately m^k times as small as in (1). Adding together the results of integration and the remainders, we find that the error of the original integral is approximately m^{k-1} times as small as in the case when we apply the chosen quadrature formula to the whole original interval.

This method is general enough and may be realized for any quadrature formula.

The rectangular formula (4) with remainder (8) considered in 8.5 is the simplest example of a compounded quadrature formula.

We shall discuss compounded formulas which are most frequently used and which are based on the simplest quadratures presented earlier.

Compounded trapezoid formula. Assume that we have to calculate the integral of the function f on the interval $[a, b]$. We divide the interval $[a, b]$ into m equal subintervals with boundary points $a = x_0, x_1, \ldots, x_m = b$

so that the length of each subinterval is $h = (b-a)/m$. We represent the integral as the sum

$$\int_a^b f(x)\,dx = \int_{x_0}^{x_1} f(x)\,dx + \int_{x_1}^{x_2} f(x)\,dx + \ldots + \int_{x_{m-1}}^{x_m} f(x)\,dx$$

and to each term on the right-hand side of this relation apply the trapezoid formula with a remainder [see relations (11) and (12) in 8.5]. Collecting terms, we obtain

$$\int_a^b f(x)\,dx = \frac{b-a}{m}\left(\frac{f_0+f_m}{2} + \sum_{i=1}^{m-1} f_i\right) - \frac{(b-a)^3}{12m^3}\sum_{i=1}^{m} f''(\eta_i), \quad \eta_i \in (x_{i-1}, x_i). \tag{2}$$

Assuming that the second derivative of the function being integrated is continuous throughout the interval $[a, b]$, we have, by virtue of Weierstrass' theorem,

$$\sum_{i=1}^{m} f''(\eta_i) = mf''(\eta), \quad \eta \in (a, b).$$

Substituting this expression for the sum into relation (2), we get a *compounded trapezoid formula*

$$I \equiv \int_a^b f(x)\,dx \cong \frac{b-a}{m}\left(\frac{f_0+f_m}{2} + \sum_{i=1}^{m-1} f_i\right) \equiv I_2^m \tag{3}$$

with a remainder

$$R_2^m[f] = -\frac{(b-a)^3}{12m^2} f''(\eta). \tag{4}$$

Consequently, we can represent the estimate of the error of quadrature (3) in the form

$$\Delta_1 = |I - I_2^m| \leqslant \frac{(b-a)^3}{12m^2} M_2, \tag{5}$$

where $M_2 = \max\limits_{[a,\,b]} |f''(x)|$. The estimate obtained cannot be improved since it is attained, say, on the parabola $f = x^2$, which fact can be proved by a direct verification. Note that we have performed similar calculations when we analysed the quadrature formula (4) in 8.4. See how this can be done for the compounded formula (3).

We calculate the left-hand side of the quadrature formula (3) for $f = x^2$:

$$I = \int_a^b x^2 \, dx = \frac{b^3 - a^3}{3} .$$

We calculate the sum:

$$I_2^m = \frac{b-a}{m} \left[\frac{a^2+b^2}{2} + \sum_{i=1}^{m-1} \left(a + i \frac{b-a}{m} \right)^2 \right].$$

Transforming the expression in brackets and using the following formulas for summation:

$$\sum_{i=1}^{m-1} i = \frac{(m-1)\,m}{2} \quad \sum_{i=1}^{m-1} i^2 = \frac{(m-1)\,m\,(2m-1)}{6} ,$$

we get the required quantity

$$I_2^m = \frac{b^3 - a^3}{3} + \frac{(b-a)^3}{6m^2} .$$

Thus the error of the quadrature formula (3) for the function $f = x^2$ is

$$\Delta_1 = |I - I_2^m| = \frac{(b-a)^3}{6m^2} ,$$

and this exactly coincides with the right-hand side of estimate (5) since $M_2 = 2$.

Since the operations $I[f]$ and $I_2^m[f]$ are linear and formula (3) is accurate for any linear function, we can assert that estimate (5) is attained on a second-degree arbitrary parabola.

Expression (5) for the error Δ_1 of the method of the quadrature formula (3) shows that $\lim_{m \to \infty} \Delta_1 = 0$, i.e. increasing m, we can ensure any preassigned accuracy I_2^m in the sense of the error of the method (when there is no computing error). In such cases we say that the method is *convergent*. It remains to find out how the error of the initial data (the values f_i of the function) affects the error of the result. Since the right-hand side of the quadrature formula (3) is linear with respect to the values of the function being integrated, the total computing error is proportional to the error of calculation of each value of the function (provided that they are equal). We can therefore expect that when the initial data are sufficiently accurate, the total computing error will be sufficiently small.

We shall derive formulas for the number of divisions m and the permissible error of each value f_i of the function, which ensure the required accuracy ε when use is made of the quadrature formula (3) with error (5).

First of all, using the algorithm of solving Problem II given in 8.3, we represent ε as the sum $\varepsilon = \varepsilon_1 + \varepsilon_2 + \varepsilon_3$ (if, for instance, $\varepsilon = 10^{-k}$, then we usually set $\varepsilon_1 = 0.3\cdot 10^{-k}$, $\varepsilon_2 = 0.2\cdot 10^{-k}$, $\varepsilon_3 = 0.5\cdot 10^{-k}$).

Next we choose the number of divisions m for which the inequality $\Delta_1 \leqslant \varepsilon_1$ will be satisfied. For this purpose, by virtue of (5), it is sufficient to require that

$$\frac{(b-a)^3}{12m^2} M_2 \leqslant \varepsilon_1.$$

Transforming this inequality, we get the following formula for the number of divisions:

$$m \geqslant (b-a)\sqrt{\frac{(b-a)M_2}{12\varepsilon_1}}. \qquad (6)$$

Thus the error of the method Δ_1 does not exceed ε_1 if the number of divisions m satisfies inequality (6).

Let us find out what must be the error of the values of the function being integrated for the total error of calculating I_2^m from formula (3) not to exceed ε_2. Let the required error be $\Delta[f_i]$. Then, using formula (3), we obtain

$$\Delta[I_2^m] \leqslant \frac{b-a}{m}(\Delta[f_i] + (m-1)\Delta[f_i]) = (b-a)\,\Delta[f_i].$$

Consequently, for the inequality $\Delta[I_2^m] \leqslant \varepsilon_2$ to hold true, it is sufficient to require that

$$\Delta[f_i] \leqslant \varepsilon_2/(b-a). \qquad (7)$$

Example 1. Using the compounded trapezoid formula, calculate the integral $\int_0^1 \frac{dx}{1+x}$ with an accuracy of 0.01.

△ Let $\varepsilon_1=0.004$, $\varepsilon_2=0.001$, $\varepsilon_3=0.005$. We find that $M_2=\max_{[0,1]}\left|\left(\frac{1}{1+x}\right)''\right|=2$. Employing inequality (6), we find the number of divisions we need:

$$m\geqslant(1-0)\cdot\sqrt{\frac{(1-0)\cdot 2}{12\cdot 0.004}}=6.4$$

We assume that $m=7$. Inequality (7) yields a value $\Delta[f_i]=0.001$ for the permissible error of the values of the integrand. We compile a table of the necessary values of the function being integrated with three valid digits:

x	0	1/7	2/7	3/7	4/7	5/7	6/7	1
$(1+x)^{-1}$	1	0.875	0.778	0.700	0.636	0.583	0.538	0.500

From formula (3) we obtain

$$I_7^2=\frac{1-0}{7}\left(\frac{1+0.5}{2}+0.875+0.778+0.700+0.636\right.$$
$$\left.+0.583+0.538\right)=0.694.$$

Rounding off the result obtained, we finally have

$$\int_0^1 \frac{dx}{1+x}=0.69\pm 0.01.$$

Compare the solution we have obtained with that of Example 6 from 8.4 and Example 3 from 8.5. ▲

Simpson's compounded formula. In this case we divide the integration interval into an even number $2m$ of equal subintervals with boundary points $a=x_0, x_1, \ldots, x_{2m}=b$ so that the length of each subinterval is $h=(b-a)/(2m)$. We represent the integral as the sum

$$\int_a^b f(x)\,dx=\int_{x_0}^{x_2} f(x)\,dx+\int_{x_2}^{x_4} f(x)\,dx+\ldots+\int_{x_{2m-2}}^{x_{2m}} f(x)\,dx.$$

We apply Simpson's formula with a remainder [see relations (11) and (12) from 8.5] to each term on the right-hand side of this relation. Collecting terms, we obtain

$$\int_a^b f(x)\,dx = \frac{b-a}{6m}(f_0 + f_{2m} + 4\sigma_1 + 2\sigma_2)$$

$$-\frac{(b-a)^5}{2880m^5}\sum_{i=1}^{m} f^{\mathrm{IV}}(\eta_i), \tag{8}$$

$$\eta_i \in (x_{2i-2},\ x_{2i});$$

$$\sigma_1 = f_1 + f_3 + \ldots + f_{2m-1},\ \sigma_2 = f_2 + f_1 + \ldots + f_{2m-2}.$$

Assuming that the fourth derivative of the function being integrated is continuous on the whole interval $[a,\ b]$, we have, by virtue of Weierstrass' theorem a relation

$$\sum_{i=1}^{m} f^{\mathrm{IV}}(\eta_i) = mf^{\mathrm{IV}}(\eta),\quad \eta \in (a,\ b).$$

Substituting the expression obtained for the sum of the derivatives into relation (8), we arrive at *Simpson's compounded formula*

$$I \equiv \int_a^b f(x)\,dx \cong \frac{b-a}{6m}(f_0 + f_{2m} + 4\sigma_1 + 2\sigma_2) \equiv I_3^m, \tag{9}$$

$$\sigma_1 = f_1 + f_3 + \ldots + f_{2m-1},\ \sigma_2 = f_2 + f_1 + \ldots + f_{2m-2}$$

with a remainder

$$R_3^m[f] = -\frac{(b-a)^5}{2880m^4} f^{\mathrm{IV}}(\eta),\ \eta \in (a,\ b). \tag{10}$$

Consequently, we can represent the estimate of the error of the quadrature formula (9) as

$$\Delta_1 = |I - I_3^m| \leqslant \frac{(b-a)^5}{2880m^4} M_4, \tag{11}$$

where $M_4 = \max\limits_{[a,b]} |f^{\mathrm{IV}}(x)|$. This estimate cannot be improved since it is attained, say, on the function $f = x^4$.

By analogy with the trapezoid formula, we shall get specific formula for the number of divisions m and the permissible error for each value f_i which ensures the required accuracy of ε when formula (9) is used with error (11).

Let $\varepsilon = \varepsilon_1 + \varepsilon_2 + \varepsilon_3$. Taking relation (11) into account, we require, for the inequality $\Delta_1 \leqslant \varepsilon_1$, that

$$\frac{(b-a)^5}{2880m^4} M_4 \leqslant \varepsilon_1.$$

From this we find the condition for the number of divisions $2m$ for which the error of the method does not exceed ε_1:

$$2m \geqslant (b-a)\sqrt[4]{\frac{(b-a)M_4}{180\varepsilon_1}}. \qquad (12)$$

We shall find out now what the error of the values of the integrand should be for the total error of calculating I_3^m from formula (9) not to exceed ε_2. Let the required error be $\Delta[f_i]$. Then, using formula (9), we have

$$\Delta[I_3^m] \leqslant \frac{b-a}{6m}(2\Delta[f_i] + 4m\Delta[f_i] + 2(m-1)\Delta[f_i]) = (b-a)\Delta[f_i].$$

Consequently, for the inequality $\Delta[I_3^m] \leqslant \varepsilon_2$ to hold true, it is sufficient to require that

$$\Delta[f_i] \leqslant \varepsilon_2/(b-a). \qquad (13)$$

As could be expected, we have got the same result as for the trapezoid formula. This is because both formulas (3) and (9) are of the same kind (3) from 8.3, all the coefficients A_i being positive and their sum equal to unity.

■ **Example 2.** Using Simpson's compounded formula, calculate the integral $\int_0^1 \frac{dx}{1+x}$ with an accuracy of 0.001.

△ Let $\varepsilon_1 = 0.00045$, $\varepsilon_2 = 0.00005$, $\varepsilon_3 = 0.00005$. We find that $M_4 = \max\limits_{[0,1]} \left|\left(\frac{1}{1+x}\right)^{\mathrm{IV}}\right| = 24$. With the aid of inequalities (12), we find the number of divisions we must make:

$$2m \geqslant (1-0)\cdot\sqrt[4]{\frac{(1-0)\cdot 24}{180\cdot 0.00045}} = 4.1 \ldots$$

Since $2m$ is an even number, we assume that $2m = 6$. Inequality (13) for the permissible error of the values of the integrand yields a value $\Delta[f_i] = 0.00005$. We compile a table of necessary values of the integrand with four valid digits:

x	0	1/6	2/6	3/6	4/6	5/6	1
$(1+x)^{-1}$	1	0.8571	0.75	0.6667	0.6	0.5455	0.5

From formula (9) we find that

$$I_3^6 = \frac{1-0}{18}[1+0.5+4\cdot(0.8571+0.6667+0.5455)+2\cdot(0.75+0.6)] = 0.69317\ \ldots$$

Rounding off the result, we finally have

$$\int_0^1 \frac{dx}{1+x} = 0.693 \pm 0.001.\ \blacktriangle$$

Let us analyse the result obtained in Example 2. Comparing the approximate value $I_3^6 = 0.69317 \ldots$ with the exact value of the integral $I = 0.69314 \ldots$, we note that the real difference of the exact and the approximate value (about 0.00003) is approximately 17 times smaller than the preassigned permissible error ($\varepsilon_1 + \varepsilon_2 = 0.0005$). A natural question arises: why is the difference of the theoretically predicted and the practically obtained error so great?

Before answering this question, we point out that such a difference between the true error and the estimate required can be observed not very frequently and can be predicted in cases similar to that considered in Example 2. And what is more, it can be reduced to a minimum by taking into account some auxiliary considerations and constructing more cleverly the computation process.

Indeed, in Example 2, by virtue of inequality (12), the error of the method 0.0004 is associated with the number of divisions $2m \geqslant 4.1 \ldots$. We could not choose the limiting value $4.1 \ldots$ since $2m$ must be integral and even and the least permissible value $2m = 6$ we have chosen is naturally associated with a smaller error of the method, approximately equal to 0.0001. This is the first cause for a decrease in the practical error.

Next, although all the estimates of the error were the least and could not be improved, they were attained on polynomials whose derivative appearing as a factor in the expression for the remainder was a constant quantity. Now if this derivative varies considerably on the interval of integration (and this is quite natural when the interval is large), then the corresponding estimate proves to be far from optimal, as a rule. If this is the case, it is expedient to determine the number of divisions proceeding from the form of the remainder given in relations (2) and (8). For Simpson's formula, for instance, we would get the following inequality for $2m$:

$$\frac{(b-a)^5}{180(2m)^4}\cdot\frac{1}{m}\cdot\sum_{i=1}^{m} M_4^i \leqslant \varepsilon_1, \tag{14}$$

where $M_4^i = \max\limits_{[x_{2i-2},\, x_{2i}]} |f^{\mathrm{IV}}(x)|$. Evidently, for great variations of the fourth derivative on the integration interval, inequality (14) is $\sqrt[4]{\dfrac{M_4}{(1/m)\sum\limits_{i=1}^{m} M_4^i}}$ times as advantageous as inequality (12) as concerns the number of divisions. In principle, we see under the sign of the root the ratio between the maximum absolute value of the fourth derivative throughout the integration interval and its mean value calculated for m intervals $[x_{2i-2}, x_{2i}]$, $i = 1, 2, \ldots, m$. Thus the use of inequality (14) in Example 2 for the number of divisions yields a value $2m = 4$ and for $2m = 6$ the estimate of the error of the method comes down to 0.000045, and this is only one and a half times larger than the true error. This is a good result.

Exercises

1. Calculate the definite integral $\int\limits_2^8 \sqrt{x+2}\,\mathrm{d}x$ using the formula for left rectangles for $n = 6$.

2. Using the formula for right rectangles for $n = 8$, calculate $\int\limits_1^9 \frac{\mathrm{d}x}{1+x}$.

3. Using the trapezoid formula for $n = 8$, calculate $\int\limits_0^8 \frac{\mathrm{d}x}{1+x}$.

4. Using the trapezoid formula, calculate $\int\limits_0^5 \frac{\mathrm{d}x}{\sqrt{x^2+4}}$ setting $n = 5$.

5. From Simpson's formula calculate $\int\limits_0^1 \frac{\mathrm{d}x}{x^2+9}$ setting $2m = 10$.

6. Calculate $\int\limits_0^1 \frac{\cos x}{1+x}\,\mathrm{d}x$ using Simpson's formula. Set $2m = 10$.

7. Calculate the integral $\int\limits_{-1}^{1} \frac{dx}{x+3}$ using the Gaussian quadrature formula for $2m = 6$.

8. The function is tabulated as follows:

x	0.525	0.526	0.527	0.528
f	0.50121	0.50208	0.50294	0.50381

Using the method of numerical differentiation, find the first derivative at the point $x^* = 0.525$.

9. Find the first derivative at the point $x^* = 50$ for a function given as the following table:

x	50	55	60	65
f	1.6990	1.7404	1.7782	1.8129

10. Calculate the integral $\int\limits_{0}^{1} f(x)\,dx$ with an accuracy of ε:

(a) $f(x) = x^3 \cos x$, $\varepsilon = 0.001$, (b) $f(x) = x^2 \sin x$, $\varepsilon = 0.001$, (c) $f(x) = xe^x$, $\varepsilon = 0.0001$, (d) $f(x) = x\sqrt{x}e^x$, $\varepsilon = 0.0001$, (e) $f(x) = e^{x^2}$, $\varepsilon = 0.0001$, (f) $f(x) = e^{x\sqrt{x}}$, $\varepsilon = 0.0001$, (g) $f(x) = x^2 + 1$, $\varepsilon = 0.0001$, (h) $f(x) = \sin x$, $\varepsilon = 0.001$, (i) $f(x) = e^x$, $\varepsilon = 0.001$.

11. Calculate the values of the derivative of the function $f(x)$ at the point $x = x_1$ using the four-digit tables of trigonometric functions with a stepsize of 1°. Find the absolute value of the result.

(a) $f(x) = \sin x$, $x_1 = 41°$, (b) $f(x) = \cos x$, $x_1 = 27.5°$, (c) $f(x) = \tan x$, $x_1 = 50°$, (d) $f(x) = \sin x$, $x_1 = 17°30'$, (e) $f(x) = \cos x$, $x_1 = 63°$, (f) $f(x) = \tan x$, $x_1 = 33.5°$.

Chapter 9

Approximate Solutions of Ordinary Differential Equations

9.1. Differential Equations

An equation in which the unknown function is under the sign of the derivative or differential is a *differential equation*, for example

$$\frac{dy}{dx}=2(y-3),\quad \frac{d^2y}{dt^2}=t+1,\quad \frac{\partial^2 z}{\partial x^2}+\frac{\partial^2 z}{\partial y^2}=0,$$

$$y'=x^2,\quad x\,dy=y^3\,dx.$$

If the unknown function entering into a differential equation depends only on one independent variable, the differential equation is *ordinary*. The differential equations

$$x^2\cdot\frac{d^2y}{dx^2}=2,\quad 2s\,dt=t\,ds$$

are of this kind.

Now if the unknown function entering into a differential equation is a function of two or more independent variables, then it is a *partial differential equation*. For instance, the equation

$$\frac{\partial^2 z}{\partial x^2}+\frac{\partial^2 z}{\partial y^2}=0$$

is a partial differential equation.

The *order* of a differential equation is the highest order of the derivative (or of the differential) entering into the equation. For instance, the equations

$$\frac{d^2s}{dt^2}=t-1,\quad \frac{\partial^2 z}{\partial x^2}+\frac{\partial^2 z}{\partial y^2}=1$$

are of the second order and the equations

$$\frac{ds}{dt}\cdot\cos t+\sin t=1,\quad (x^2-y^2)\,dx+(x^2+y^2)\,dy=0$$

are of the first order.

In this chapter we study only ordinary differential equations.

In the most general case an ordinary differential equation of order n contains an independent variable, an unknown function and its derivatives or differentials up to order n inclusive and has the form

$$F(x, y, y', y'', \ldots, y^{(n)}) = 0. \quad (1)$$

In this equation x is an independent variable, y is an unknown function, y', y'', . . ., $y^{(n)}$ are derivatives of this function. .

If the left-hand side of the differential equation (1) is a polynomial with respect to the derivative of the unknown function, then the degree of this polynomial is the *degree* of the differential equation.

For instance, the equation

$$(y'')^4 + (y')^2 - y^6 + x^7 = 0$$

is a second-order equation of the fourth degree and the equation

$$(y')^2 + x^4y^5 - y^8 + x^{10} = 0$$

is a first-order equation of the second degree.

An nth-order differential equation resolved for the highest derivative can be written as

$$y^{(n)} = f(x, y, y', y'', \ldots, y^{(n-1)}). \quad (2)$$

The *solution* (or *integral*) of equation (2) is any differentiable function $y = \varphi(x)$ which satisfies this equation, i.e. such that being substituted into equation (2), it turns the equation into an identity.

The graph of the solution of an ordinary differential equation is known as the *integral curve* of this equation.

The solution of a differential equation which contains as many independent arbitrary constants (parameters) as its order is a *general solution* (or *general integral*) of this equation.

In terms of geometry, the general solution of a differential equation is a family of integral curves of that equation.

A *particular solution* of a differential equation is any solution which can be selected from the general solution by giving a definite set of values to the arbitrary constants.

The arbitrary constants entering into the general solution are found from the so-called initial conditions.

A problem with initial conditions is posed as follows: find the solution $y = \varphi(x)$ of the equation $y^{(n)} = f(x, y, y', y'', \dots, y^{(n-1)})$ which satisfies the auxiliary conditions that the solution $y = \varphi(x)$ must assume, together with its derivatives up to the order $n - 1$, the preassigned numerical values $y_0, y'_0, y''_0, \dots, y_0^{(n-1)}$ for the given numerical value $x = x_0$ of the independent variable x:

$$y = y_0,\ y' = y'_0,\ y'' = y''_0,\ \dots,\ y^{(n-1)} = y_0^{(n-1)} \text{ for } x = x_0. \tag{3}$$

Conditions (3) are the *initial conditions* or *values*, the numbers $x_0, y_0, y'_0, y''_0, \dots, y_0^{(n-1)}$ are the *initial data* of the solution, and the problem of finding the solution $y = \varphi(x)$ of the differential equation (2), satisfying the initial conditions (3), is an *initial-value problem* (or *Cauchy's problem*).

In the case of a first-order equation, i.e. when $n = 1$, we get Cauchy's problem for the equation $y' = f(x, y)$ with the initial condition $x = x_0$, $y = y_0$.

In terms of geometry, Cauchy's problem (for a first-order equation) consists in isolating, from the whole set of integral curves which constitutes the general solution, the integral curve which passes through the point M_0 with coordinates $x = x_0$, $y = y_0$.

Example. The general solution of the differential equation $\frac{dy}{dx} = 2x$, with the initial condition $y_0 = 2$ for $x_0 = 1$, has the form $y = x^2 + C$. It is a family of parabolas. If we substitute the initial data into the general solution, we get $2 = 1 + C$, i.e. $C = 1$. Consequently, the particular solution which satisfies the indicated initial condition is $y = x^2 + 1$. In terms of geometry, this means that from the whole set of parabolas which constitute the general solution of the differential equation we choose the parabola which passes through the point M_0 (1, 2) (Fig. 9.1).

Cauchy's problem has a unique solution which satisfies the condition $y(x_0) = y_0$ if the function $f(x, y)$ is continuous in a certain domain $R_{[a,b]} = \{ |x - x_0| < a, |y - y_0| < b \}$ and satisfies, in this domain, the *Lipschitz condition*

$$|f(x, \bar{y})| - |f(x, y)| \leqslant N\,|\bar{y} - y|,$$

where N is the Lipschitz constant dependent on a and b (a and b are the boundaries of the domain).

The methods of the exact integration of a differential equation suit only a small part of the equations which are encountered in practice.

Therefore of considerable significance are approximate methods of solving differential equations which can be divided into two groups according to the form of the representation of the solution:

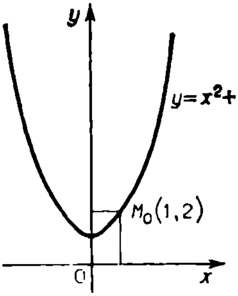


Fig. 9.1

(1) **analytic methods** which give an approximate solution of a differential equation in the form of an analytic expression,

(2) **numerical methods** which give an approximate solution in the form of a table.

For the first group of methods, we discuss in this chapter the method of successive approximations (Picard's method) and the method of integration of differential equations by means of power series, and for the second group, Euler's method and its modifications, the methods of Runge-Kutta and Adams.

9.2. The Method of Successive Approximations (Picard's Method)

This method was first used in the process of proving the theorem of the existence and uniqueness of the solution of differential equations. It is known as **Picard's method.**

Consider an equation

$$y' = f(x, y) \tag{1}$$

whose right-hand side is continuous in the rectangle $\{ |x - x_0| \leqslant a, \ |y - y_0| \leqslant b\}$ and has a continuous partial derivative with respect to y. We have to find a solution of equation (1) which satisfies the initial condition

$$x = x_0, \ y(x_0) = y_0. \tag{2}$$

Integrating both sides of the equation from x_0 to x, we obtain

$$\int_{y_0}^{y} dy = \int_{x_0}^{x} f(x, y)\, dx,$$

or

$$y(x) = y_0 + \int_{x_0}^{x} f(x, y)\, dx. \tag{3}$$

Equation (1) is replaced by the integral equation (3) in which the unknown function y is under the integral sign. The integral equation (3) satisfies the differential equation (1) and the initial condition (2). Indeed,

$$y(x_0) = y_0 + \int_{x_0}^{x} f(x, y)\, dx = y_0.$$

Replacing the function y in relation (3) by its value y_0, we get the first approximation

$$y_1(x) = y_0 + \int_{x_0}^{x} f(x, y_0)\, dx.$$

Then we replace y in relation (3) by the value y_1 already found and get the second approximation

$$y_2(x) = y_0 + \int_{x_0}^{x} f(x, y_1)\, dx.$$

Continuing the process, we find in succession

$$y_3(x) = y_0 + \int_{x_0}^{x} f(x,\ y_2)\,dx,$$

$$\dots\dots\dots\dots\dots$$

$$y_n(x) = y_0 + \int_{x_0}^{x} f(x,\ y_{n-1})\,dx.$$

Thus we form a sequence of functions

$$y_1(x),\ y_2(x),\ y_3(x),\ \dots,\ y_n(x).$$

The following theorem, which we give without proof, is valid.

Theorem. *Assume that the function* $f(x,\ y)$ *is continuous and has a bounded partial derivative* $f'_y(x,\ y)$ *in the neighbourhood of the point* $(x_0,\ y_0)$. *Then, in a certain interval, which contains the point* x_0, *the sequence* $\{y_i(x)\}$ *converges to the function* $y(x)$ *which is a solution of the differential equation* $y = f(x,\ y)$ *and satisfies the condition* $y(x_0) = y_0$.

The estimate of the error of Picard's method can be found from the formula

$$|y - y_n| \leqslant N^n M \frac{h^{n+1}}{(n+1)!}, \tag{4}$$

where $M = \max|f(x,\ y)|$ for $x,\ y \in R_{[a,b]}$ and N is the Lipschitz constant for the domain $R_{[a,b]}$ equal to $N = \max|f'_y(x,\ y)|$. The quantity h for determining the domain $[x_0 - h \leqslant x \leqslant x_0 + h]$ can be calculated from the formula

$$h = \min(a,\ b/M), \tag{5}$$

where a and b are the boundaries of the domain R.

Example. Use Picard's method to solve the differential equation $y' = x^2 + y^2$ satisfying the initial conditon $x_0 = 0,\ y(x_0) = y_0 = 0$.

△ We pass to the integral equation

$$y = y_0 + \int_{x_0}^{x} (x^2 + y^2)\,dx,$$

or, with due account for the initial conditions, to

$$y=\int_0^x (x^2+y^2)\,dx.$$

We get the successive approximations

$$y_1=\int_0^x (x^2+y_0^2)\,dx=\int_0^x (x^2+0)\,dx=\frac{x^3}{3},$$

$$y_2=\int_0^x (x^2+y_1^2)\,dx=\int_0^x \left(x^2+\frac{x^6}{9}\right)dx=\frac{x^3}{3}+\frac{x^7}{63},$$

$$y_3=\int_0^x (x^2+y_2^2)\,dx=\int_0^x \left(x^2+\frac{x^6}{9}+\frac{2x^{10}}{189}+\frac{x^{14}}{3969}\right)dx$$

$$=\frac{x^3}{3}+\frac{x^7}{63}+\frac{2x^{11}}{2079}+\frac{x^{15}}{59\,535}.$$

We use formula (4) to estimate the error of the third approximation:

$$|y-y_n|\leqslant N^nM\frac{h^{n+1}}{(n+1)!}.$$

Since the function $y'=x^2+y^2$ is defined and continuous on the entire plane, we can take any numbers as a and b. For definiteness we choose a rectangle

$$R\ \{|x-x_0|\leqslant 0.5,\ |y-y_0|\leqslant 1\},$$

i.e. $R\ \{-0.5\leqslant x\leqslant 0.5,\ -1\leqslant y\leqslant 1\}$.
Then

$$M=\max|f(x,\ y)|=\max(x^2+y^2)=1.25,$$
$$N=\max|f'_y(x,\ y)|=\max|2y|=2.$$

Since $a=0.5$, $b/M=0.8$, we have, in accordance with formula (5),

$$h=\min(a,\ b/M)=0.5.$$

The solution y will be defined for $-0.5\leqslant x\leqslant 0.5$. For $n=3$ we find that

$$|y-y_3|\leqslant 1.25\cdot 2^3\cdot 0.5^4/4!=5/192.$$

We have got a very rough estimate of the error. in actual fact the error is considerably smaller. ▲

9.3. Integrating Differential Equations by Means of Power Series

The method of successive differentiation. Consider an nth-order differential equation

$$y^{(n)} = f(x, y, y', \ldots, y^{(n-1)}) \tag{1}$$

with the initial conditions

$$x = x_0, \ y(x_0) = y_0, \ y'(x_0) = y'_0, \ \ldots, \ y^{(n-1)}(x_0) = y_0^{(n-1)}. \tag{2}$$

The right-hand side of this equation is an analytic function at the initial point $M_0(x_0, y_0, y'_0, \ldots, y_0^{(n-1)})$. We represent the solution $y = y(x)$ of equation (1) in the neighbourhood of the point x_0 as a Taylor series

$$y = y_0 + y'_0(x - x_0) + \frac{y''_0}{2!}(x - x_0)^2 + \cdots + \frac{y_0^{(n)}}{n!}(x - x_0)^n + \ldots, \tag{3}$$

where $|x - x_0| < h$ and h is a sufficiently small quantity. To find the coefficients of series (3), we differentiate equation (1) with respect to x as many times as necessary, using conditions (2).

In practical calculations, the quantity $|x - x_0|$ is taken so small that for the required degree of accuracy the remainder of the series can be neglected.

If $x_0 = 0$, we get a Taylor series in the powers of x:

$$y = y_0 + y'_0 x + \frac{y''_0}{2!}x^2 + \ldots + \frac{y_0^{(n)}}{n!}x^{(n)} + \ldots. \tag{4}$$

Example 1. Find a solution of the differential equation $y = y - 4x + 3$ which would satisfy the initial condition $x_0 = 0$, $y_0 = 3$.

△ Substituting $y_0 = 3$ into expansion (4), we obtain

$$y = 3 + \frac{y'_0}{1}x + \frac{y''_0}{2!}x^2 + \frac{y'''}{3!}x^3 + \ldots + \frac{y_0^{(n)}}{n!}x^{(n)} + \ldots. \tag{*}$$

Successively differentiating this equation, we have

$y'' = y' - 4 = y - 4x - 1$, $y''' = y'' = y - 4x - 1$, $y^{\mathrm{IV}} = y^{\mathrm{III}}$ and so on.

Using the initial condition, we find that

$$y'_0 = y_0 - 4x_0 + 3 = 3 + 3 = 6,\ y'' = y_0 - 4x_0 - 1 = 3 - 1 = 2,$$

$$y'''_0 = 2,\ y_0^{\mathrm{IV}} = 2,\ \ldots,\ y_0^{(n)} = 2.$$

Substituting y'_0, y''_0, y'''_0 into the right-hand side of relation (*), we obtain

$$y = 3 + 6x + x^2 + \frac{x^3}{3} + \frac{x^4}{12} + \ldots .$$

The function

$$y = 2e^x + 4x + 1$$

is the exact soultion of the given equation.

If we set $h = 0.1$, we can compile a table of values of the solution of the given differential equation. ▲

Table 9.1

x_i	0	0.1	0.2	0.3
The values obtained from the analytic solution	3	3.6021	4.2428	4.8996
An approximate solution by means of a power series	3	3.6103	4.2427	4.8999

Example 2. Find a solution of the differential equation $y'' - x^2y = 0$ which would satisfy the initial condition $x_0 = 0$, $y_0 = 1$, $y'_0 = 0$.

△ As in Example 1, we shall seek the solution of this equation as a series (4) in the powers of x. Since $y_0 = 1$, $y'_0 = 0$, series (4) has the form

$$y = 1 + \frac{y''_0}{2!}x^2 + \frac{y'''_0}{3!}x^3 + \ldots + \frac{y_0^{(n)}}{n!}x^n + \ldots . \qquad (*)$$

We rewrite the given differential equation as $y'' = x^2y$. Successively differentiating this relation, we have

$$\begin{aligned} y''' &= 2xy + x^2y', \\ y^{\mathrm{IV}} &= 2y + 2xy' + 2xy' + x^2y'' = 2y + 4xy' + x^2y'', \\ y^{\mathrm{V}} &= 2y' + 4y' + 4xy'' + 2xy'' + x^2y''' = 6y' + 6xy'' + x^2y''', \\ y^{\mathrm{VI}} &= 12y'' + 8xy''' + x^2y^{\mathrm{IV}}, \\ y^{\mathrm{VII}} &= 20y''' + 10xy^{\mathrm{IV}} + x^2y^{\mathrm{V}}, \\ y^{\mathrm{VIII}} &= 30y^{\mathrm{IV}} + 12xy^{\mathrm{V}} + x^2y^{\mathrm{VI}}. \end{aligned}$$

Substituting successively the initial conditions $x_0 = 0$, $y_0 = 1$, $y'_0 = 0$ into each of the relations obtained, we get $y''(0) = 0$, $y^{\mathrm{III}}(0) = 0$, $y^{\mathrm{IV}}(0) = 2$, $y^{\mathrm{V}} = y^{\mathrm{VI}} = y^{\mathrm{VII}} = 0$, $y^{\mathrm{VIII}} = 30 \cdot 2 = 60$.

Substituting finally y'', y''', y^{IV}. . . into the right-hand side of relation (*), we get the required solution

$$y = 1 + \frac{1}{12} x^4 + \frac{1}{672} x^8 + \ldots . \ \blacktriangle$$

The method of expanding a solution in a series can also be used to solve systems of ordinary differential equations.

Example 3. Find a solution of the system

$$\begin{cases} \dfrac{dx}{dt} = x \cos t - y \sin t, \\ \dfrac{dy}{dt} = x \sin t + y \cos t, \end{cases}$$

which would satisfy the initial conditions $x(0) = 1$, $y(0) = 0$.

$\triangle$ We set

$$x = x(0) + x'(0)\, t + x''(0) \frac{t^2}{2} + \frac{x'''(0)}{3!} t^3 + \ldots,$$

$$y = y(0) + y'(0)\, t + y''(0) \frac{t^2}{2} + \frac{y'''(0)}{3!} t^3 + \ldots .$$

Differentiating the equations of the system, we obtain

$$x'' = x' \cos t - x \sin t - y' \sin t - y \cos t,$$

$$y'' = x' \sin t + x \cos t + y' \cos t - y \sin t \text{ and so on.}$$

Substituting in succession the initial conditions into each of the relations obtained, we find that $x'(0) = \cos t\,|_{t=0} = 1$, $y'(0) = 0$, $x''(0) = 1$, $y''(0) = 1$, $x'''(0) = 0$, $y'''(0) = 3$ and so on.

The required solution has the form

$$x = 1 + t + \frac{t^2}{2} + \ldots, \quad y = \frac{t^2}{2} + \frac{t^3}{2} + \ldots . \ \blacktriangle$$

The method of undetermined coefficients. Consider a differential equation

$$y' = f(x, y) \tag{5}$$

with the initial condition $y(x_0) = y_0$. The method of undetermined coefficients consists in seeking a solution of equation (5) in the form of a series with unknown coefficients

$$y = a_0 + a_1 (x - x_0) + a_2 (x - x_0)^2 + a_3 (x - x_0)^3 + \ldots, \tag{6}$$

which we can find by substituting series (6) into equation (5), equating the coefficients in the like powers of x and using the initial condition. Then we substitute the values of the coefficients $a_0, a_1, a_2, a_3, \ldots$ into series (6).

Example 4. Using the method of undetermined coefficients, find a solution of the differential equation $y' = x^2 + y^2$ which would satisfy the initial condition $x_0 = 0$, $y(x_0) = 1$.

△ Since $x_0 = 0$, series (6) assumes the form

$$y = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + \ldots. \qquad (*)$$

Substituting $x = x_0$ and $y(x_0) = 1$ into relation (*), we get $a_0 = 1$.

For what follows, it is convenient to expand the right-hand side of the equation $y' = x^2 + y^2$ in the powers of $(y - 1)$:

$$y' = x^2 + [(y-1)+1]^2 = x^2 + 1 + 2(y-1) + (y-1)^2. \qquad (**)$$

Differentiating series (*), we have

$$y' = a_1 + 2a_2x + 3a_3x^2 + 4a_4x^3 + \ldots. \qquad (***)$$

We substitute now the expression for y' from relation (***), the expression for y from relation (*) and $a_0 = 1$ into relation (**). Then we obtain

$$a_1 + 2a_2x + 3a_3x^2 + 4a_4x^3 + \ldots = 1 + x^2 + 2(a_1x + a_2x^2 + a_3x^3 + a_4x^4 + \ldots) + (a_1x + a_2x^2 + a_3x^3 + \ldots)^2.$$

We remove brackets on the right-hand side of the last relation, collect terms and equate the coefficients in the like powers of x. As a result we have

$$a_1 = 1,\ 2a_2 = 2a_1,\ 3a_3 = 1 + 2a_2 + a_1^2,\ 4a_4 = 2a_3 + 2a_1a_2,$$

whence we have $a_1 = 1$, $a_2 = 1$, $a_3 = 4/3$, $a_4 = 7/6$.

Substituting the values of the coefficients we have found into series (*), we finally obtain

$$y = 1 + x + x^2 + \frac{4}{3}x^3 + \frac{7}{6}x^4 + \ldots. \ ▲$$

Example 5. Find a solution of the differential equation $y'' - x^2y = 0$ which would satisfy the initial conditions $x_0 = 0$, $y(x_0) = 1$, $y'(x_0) = 0$.

△ Since $x_0 = 0$, we shall seek the solution in the form of a series

$$y = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n + \ldots. \qquad (*)$$

Differentiating twice relation (*), we have

$$y' = a_1 + 2a_2x + 3a_3x^2 + 4a_4x^3 + \ldots + na_nx^{n-1}, \qquad (**)$$

$$y'' = 2a_2 + 6a_3x + 12a_4x^2 + \ldots + a_nn(n-1)x^{n-2}. \qquad (***)$$

Using the initial conditions, we find, from relations (*) and (**), that $a_0 = 1$ and $a_1 = 0$. Substituting the values of the coefficients we have found into series (*), we get

$$y = 1 + a_2x^2 + a_3x^3 + a_4x^4 + \ldots + a_nx^n. \qquad (****)$$

To find the other coefficients of series (****), we substitute the expression for y'' from relation (***) and that for y from relation (****) into the given differential equation:

$$2a_2 + 6a_3x + 12a_4x^2 + 20a_5x^3 + 30a_6x^4 + \dots + n(n-1)a_nx^{n-2} - x^2(1 + a_2x^2 + a_3x^3 + a_4x^4 + \dots + a_nx^n + \dots) = 0.$$

Grouping terms with the same powers, we have

$$2a_2 + 6a_3x + (12a_4 - 1)x^2 + (20a_5)x^3 + (30a_6 - a_2)x^4 + (42a_7 - a_3)x^5 + \dots = 0.$$

The relation obtained is an identity. Since we seek a solution for $x \neq 0$, it remains to set all coefficients in the powers of x equal to zero, i. e.

$$a_2 = 0,\ a_3 = 0,\ 12a_4 - 1 = 0, \text{ whence } a_4 = 1/12,$$
$$a_5 = 0,\ 30a_6 - a_2 = 0, \text{ whence } a_6 = 0, \text{ and so on.}$$

The solution sought has the form

$$y = 1 + \frac{1}{12}x^4 + \frac{1}{672}x^8 + \dots\ . \ ▲$$

9.4. Numerical Integration of Differential Equations. Euler's Method

To solve ra differential equation of type $y' = f(x, y)$ by a numetical method is to find, for a given sequence of argumengs $x_0, x_1, \dots, x_n$ and the number y_0, without determining the function $y = F(x)$, the values $y_1, y_2, \dots, y_n$ such that $y_i = F(x_i)$ $(i = 1, 2, \dots, n)$ and $F(x_0) = y_0$.

Thus numerical methods make it possible, rather than finding the function $y = F(x)$, to get a table of values of that function for a given sequence of the values of the arguments. The quantity $h = x_k - x_{k-1}$ is the *step of integration*.

Let us consider some of the numerical methods.

Euler's method. This method is rather rough and is usually used for approximate calculations. However, the ideas that are at the base of Euler's method serve as starting points for other methods.

Consider a first-order differential equation

$$y' = f(x, y) \tag{1}$$

with the initial condition

$$x = x_0,\ y(x_0) = y_0. \tag{2}$$

We have to find a solution of equation (1) on the interval $[a, b]$.

We divide the interval $[a, b]$ into n equal parts and get a sequence $x_0, x_1, x_2, \ldots, x_n$, where $x_i = x_0 + ih$ $(i = 0, 1, 2, \ldots, n)$ and $h = (b - a)/n$ is the step of integration.

We choose the kth subinterval $[x_k, x_{k+1}]$ and integrate equation (1):

$$\int_{x_k}^{x_{k+1}} f(x, y)\,dx = \int_{x_k}^{x_{k+1}} y'\,dx = y(x)\Big|_{x_k}^{x_{k+1}}$$

$$= y(x_{k+1}) - y(x_k) = y_{k+1} - y_k,$$

i.e.

$$y_{k+1} = y_k + \int_{x_k}^{x_{k+1}} f(x, y)\,dx. \tag{3}$$

If in the last integral we assume the integrand function to be constant on the interval $[x_k, x_{k+1}]$ and equal to the initial value at the point $x = x_k$, then we obtain

$$\int_{x_k}^{x_{k+1}} f(x, y)\,dx = f(x_k, y_k)\cdot x\Big|_{x_k}^{x_{k+1}}$$

$$= f(x_k, y_k)(x_{k+1} - x_k) = y_k'h.$$

Then formula (3) assumes the form

$$y_{k+1} = y_k + y_k'h. \tag{3'}$$

Designating $y_{k+1} - y_k = \Delta y_k$, i.e. $y_k'h = \Delta y_k$, we get

$$y_{k+1} = y_k + \Delta y_k. \tag{4}$$

Continuing this process and assuming every time the integrand function to be constant on the corresponding subinterval and equal to its value at the initial point of the subinterval, we get a table of solutions of the differential equation on the given interval $[a, b]$.

In terms of geometry, Euler's method consists in the following. On the interval $[x_0, x_1]$ the integral curve is replaced by a segment of a tangent to it passing through the point $M_0(x_0, y_0)$. As can be seen from Fig. 9.2, the slope of this tangent is

$$\tan\alpha_0 = (y_1 - y_0)/(x_1 - x_0) = f(x_0, y_0) = y'(x_0).$$

Then a new segment of the tangent is drawn from the point $M_1(x_1, y_1)$ to the integral curve which passes through this point. The slope of this tangent is

$$\tan\alpha_1 = (y_2 - y_1)/(x_2 - x_1) = f(x_1, y_1).$$

Continuing the construction of such segments, we get *Euler's broken line*. This line passes through the given point $M_0(x_0, y_0)$ and approximates the required integral curve.

If in a certain rectangle

$$R\ \{|x - x_0| \leqslant a,\ |y - y_0| \leqslant b\}$$

the function $f(x, y)$ satisfies the condition

$$|f(x_1, y_1) - f(x_1, y_2)| \leqslant N\,|y_1 - y_2| \qquad (N = \text{const}) \tag{5}$$

and, in addition,

$$\frac{df}{dx} \qquad \frac{\partial f}{\partial x} + f\,\frac{\partial f}{\partial y}\Big| \leqslant M\ (M = \text{const}), \tag{6}$$

then we have the following estimate of the error:

$$|y(x_n) - y_n| \leqslant \frac{hM}{2N}\,[(1 + hN)^n - 1], \tag{7}$$

where $y(x_n)$ is the value of the exact solution of equation (1) for $x = x_n$ and y_n is an approximate value obtained at the nth step.

Formula (7) is mainly used in theoretical calculations. In practice use is usually made of "duplication check". First the calculation is carried out with the stepsize h and then the step is divided and the calculation is repeated with the stepsize $h/2$. The error of the more exact value y_n^* is estimated by the formula

$$|y_n^* - y(x_n)| \cong |y_n^* - y_n|. \tag{8}$$

Example 1. Use Euler's method to integrate, on the interval [0, 1.5], the differential equation $y' = y - x$ which satisfies the initial condition $x_0 = 0$, $y_0 = 1.5$; the step $h = 0.25$. Carry out the calculations with four decimal digits.

△ To make the calculations more convenient, we compile a table.

Table 9.2

i	x_i	y_i	$y'_i = y_i - x_i$	$\Delta y_i = hy'_i$
(1)	(2)	(3)	(4)	(5)
0	0	1.5000	1.5000	0.3750
1	0.25	1.8750	1.6250	0.4062
2	0.50	2.2812	1.7812	0.4453
3	0.75	2.7265	1.9765	0.4941
4	1.00	3.2206	2.2206	0.5552
5	1.25	3.7758	2.5258	0.6314
6	1.50	4.4702		

1st step. Using the initial data, we fill in the first row in columns (2) and (3).

2nd step. From the equation $y'_i = y_i - x_i$ we find y'_i ($i = 0, 1, \ldots, 5$) in column (4).

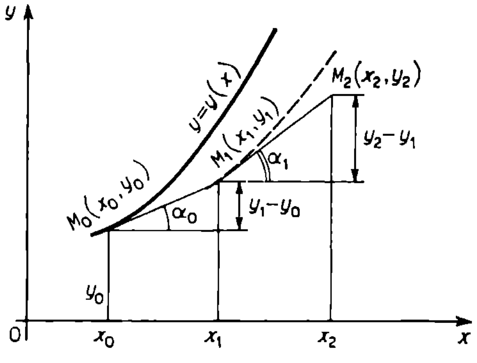


Fig. 9.2

3rd step. We multiply the contents of column (4) by h (we calculate $\Delta y_i = hy'_i$, $i = 0, 1, \ldots, 5$) and write the result in column (5) of the same row.

4th step. To the contents of column (3) we add the contents of column (5) of the same row (we calculate $y_{i+1} = y_i + \Delta y_i$, $i = 0$, 1, . . ., 5) and write the result in column (3) of the next row. We determine $x_{i+1} = x_i + h$ and then repeat steps 2, 3, and 4 until we cover the whole interval [0, 1.5]. ▲

Euler's method can be used to solve systems of differential equations and higher-order differential equations. However, in the last case the differential equations must be reduced to a system of first-order differential equations.

Consider a system of two first-order equations

$$\begin{cases} y' = f_1(x, y, z), \\ z' = f_2(x, y, z) \end{cases} \tag{9}$$

with the initial conditions

$$y(x_0) = y_0, \quad z(x_0) = z_0. \tag{10}$$

The approximate values $y(x_i) \cong y_i$ and $z(x_i) \cong z_i$ can be found from the formulas

$$y_{i+1} = y_i + \Delta y_i, \quad z_{i+1} = z_i + \Delta z_i, \tag{11}$$

where

$$\Delta y_i = hf_1(x_i, y_i, z_i) \quad \Delta z_i = hf_2(x_i, y_i, z_i)$$
$$(i = 0, 1, 2, \ldots). \tag{12}$$

Example 2. Using Euler's method, find a numerical solution of the system of differential equations $\begin{cases} y' = (z - y)x \\ z' = (z + y)x \end{cases}$ with the initial conditions $y(0) = 1.0000$, $z(0) = 1.0000$ on the interval [0, 0.6]; the step $h = 0.1$. Carry out the calculations with one extra digit.

△ To carry out he calculations, we use Table 9.3. The sequence of actions is clear from the table. ▲

Table 9.3

i	x_i	y_i	$y'_i = (z_i - y_i)x_i$	$\Delta y_i = y'_i h$	z_i	$z'_i = (z_i + y_i)x_i$	$\Delta z_i = z'_i h$
0	0	1.0000	0	0	1.0000	0	0
1	0.1	1.0000	0	0	1.0000	0.2000	0.0200
2	0.2	1.0000	0.0040	0.0004	1.0200	0.4040	0.0404
3	0.3	1.0004	0.0180	0.0018	1.0604	0.6182	0.0618
4	0.4	1.0022	0.0480	0.0048	1.1222	0.8498	0.0850
5	0.5	1.0070	0.1001	0.0100	1.2072	1.1071	0.1107
6	0.6	1.0170			1.3179		

Example 3. Using Euler's method, compile a table of values of the solution of the differential equation $y'' + \frac{y'}{x} + y = 0$ on the interval [1. 1.5] for the initial conditions $y(1) = 0.77$, $y'(1) = -0.44$; the step $h = 0.1$.

△ By means of the substitution $y' = z$, $y'' = z'$ we replace the given equation by a system of equations

$$\begin{cases} y' = z, \\ z' = -\frac{z}{x} - y \end{cases}$$

or the initial conditions $y(1) = 0.77$, $z(1) = -0.44$. We carry out the calculations with one extra digit. To make the calculations more convenient, we use Table 9.4. ▲

Table 9.4

i	x_i	y_i	$y'_i = z_i$	$\Delta y_i = hy'_i$	z_i	$z'_i = -\frac{z_i}{x_i} - y_i$	$\Delta z_i = z'_i h$
0	1.0	0.77	−0.44	−0.044	−0.44	−0.33	−0.033
1	1.1	0.726	−0.473	−0.047	−0.473	−0.296	−0.030
2	1.2	0.679	−0.503	−0.050	−0.503	−0.260	−0.026
3	1.3	0.629	−0.529	−0.053	−0.529	−0.222	−0.022
4	1.4	0.576	−0.551	−0.055	−0.551	−0.183	−0.018
5	1.5	0.521			−0.569		

9.5. Modifications of Euler's Method

Improved Euler's method. Consider a differential equation

$$y' = f(x, y) \tag{1}$$

with the initial condition

$$y(x_0) = y_0. \tag{2}$$

We have to find a solution of equation (1) on the interval $[a, b]$.

We divide the interval $[a, b]$ into n equal parts by the points $x_i = x_0 + ih$ $(i = 0, 1, 2, \ldots, n)$, where $h = (b - a)/n$ is the step of integration. The gist of the improved Euler's method is in the following. We first find auxiliary values of the required function $y_{i+1/2}$ at the points $x_{i+1/2} = x_i + \frac{h}{2}$ using the formula

$$y_{i+1/2} = y_i + \frac{h}{2} y'_i, \tag{3}$$

then find the value of the right-hand side of equation (1) at the midpoint $y'_{i+1/2} = f(x_{i+1/2}, y_{i+1/2})$ and determine

$$y_{i+1} = y_i + hy'_{i+1/2}. \tag{4}$$

Remark. We can get the estimate of the error at the point x_i by means of the "duplication check": we repeat the calculation with the step $h/2$ and approximately estimate the error of the more exact value y_i^* (for the step $h/2$) as follows:

$$|y_i^* - y(x_i)| \cong \frac{1}{3} |y_i^* - y_i|, \tag{5}$$

where $y(x)$ is the exact solution of the differential equation. This method of Euler is more accurate as compared to the method discussed in 9.4.

Example 1. Use Euler's improved method to integrate the differential equation $y' = y - x$ for the initial conditions $x_0 = 0$, $y_0 = 1.5$ on the interval $[0, 1]$ assuming $h = 0.25$. Carry out the calculations with four decimal digits.

△ We write the results of calculations in Table 9.5 which is compiled as follows.

Table 9.5

i	x_i	y_i	$y'_i = f(x_i, y_i)$	$\frac{h}{2} \cdot y'_i$	$x_{i+1/2} = x_i + \frac{h}{2}$	$y_{i+1/2} = y_i + \frac{h}{2} y'_i$	$y'_{i+1/2} = f(x_{i+1/2}, y_{i+1/2})$	$hy'_{i+1/2}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
0	0	1.5000	1.5000	0.1875	0.125	1.6875	1.5625	0.3906
1	0.25	1.8906	1.6406	0.2051	0.3750	2.0957	1.7207	0.4302
2	0.50	2.3208	1.8208	0.2276	0.6750	2.5484	1.8734	0.4684
3	0.75	2.7892	2.0392	0.2549	0.8750	3.0441	2.1691	0.5423
4	1.00	3.3315	2.3315	0.2914	1.1250	3.6229	2.4974	0.6243
5	1.25	3.9558	2.7058	0.3382	1.3750	4.2940	2.9190	0.7298
6	1.50	4.6856						

1st step. Using the initial data, we fill in the fist row in columns (2) and (3).

2nd step. From the equation $y'_i = f(x_i, y_i) = y_i - x_i$ we find y'_i for column (4) ($i = 0, 1, \ldots, 5$).

3rd step. We multiply the contents of column (4) by $h/2$ and thus determine $(h/2)\, y'_i$ and then write the result in column (5).

4th step. We get the contents of column (6) by summing up the value of x_i from column (2) and $h/2$.

5th step. We sum up the contents of column (3) with that of column (5) and write the result in column (7).

6th step. We substitute the values $x_{i+1/2}$, $y_{i+1/2}$ obtained [columns (6) and (7)] into the right-hand side of the given differential equation, determine $y'_{i+1/2}$ and write the result in column (8).

7th step. We multiply the contents of column (8) by the step of integration h and determine $hy'_{i+1/2}$ [column (9)].

8th step. We add the contents of column (3) to that of column (9) and write the result $y_{i+1} = y_i + hy'_{i+1/2}$ ($i = 0, 1, \ldots, 5$) in column (3) of the next row.

Then we repeat the whole calculation process beginning with the second step. ▲

The improved Euler-Cauchy method. The gist of this method is the following. We first find an auxiliary quantity

$$\tilde{y}_{i+1} = y_i + hy'_i, \tag{6}$$

then calculate $\tilde{y}'_{i+1} = f(x_{i+1}, \tilde{y}_{i+1})$ and find the required solution using the formula

$$y_{i+1} = y_i + h \cdot \frac{y'_i + \tilde{y}'_{i+1}}{2}. \tag{7}$$

The estimate of the error can be found from formula (5) after a repeated calculation with the stepsize $h/2$.

Example 2. Using the improved Euler-Cauchy method, integrate the differential equation from Example 1.

△ The results of calculations are given in Table 9.6. The table is filled in as follows.

1st step. Using. the initial data, we fill in the first row in columns (2) and (3).

2nd step. We determine the value of $y'_i = f(x_i, y_i) = y_i - x_i$ ($i = 0, 1, \ldots, 5$) for column (4).

3rd step. We multiply the value of y_i from column (4) by the step of integration h and write the result in column (5).

4th step. We find $x_{i+1} = x_i + h$ ($i = 0, 1, \ldots, 5$) for column (6).

5th step. We sum up the contents of column (3) with that of column (5) and write the result in column (7), i.e. find $\tilde{y}_{i+1} = y_i + hy'_i$.

Table 9.6

i	x_i	y_i	$y'_i = y_i - x_i$	hy'_i	x_{i+1}	$\tilde{y}_{i+1} = y_i + hy'_i$	$\tilde{y}'_{i+1} = \tilde{y}_{i+1} - x_{i+1}$	$h\tilde{y}'_i + 1$	$\Delta y_i = \dfrac{hy'_i + h\tilde{y}'_{i+1}}{2}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
0	0	1.5000	1.5000	0.3750	0.25	1.8750	1.625	0.4062	0.3906
1	0.25	1.8906	1.6406	0.4102	0.50	2.3008	1.8008	0.4502	0.4302
2	0.50	2.3208	1.8208	0.4552	0.75	2.7760	2.0260	0.5065	0.4808
3	0.75	2.8016	2.0516	0.5129	1.00	3.3145	2.3145	0.5786	0.5458
4	1.00	3.3474	2.3474	0.5868	1.25	3.9342	2.6842	0.6710	0.6289
5	1.25	3.9763	2.7263	0.6816	1.50	4.6579	3.1579	0.7895	0.7355
6	1.50	4.7118							

6th step. We substitute the values of x_{i+1} and y_{i+1} we have found into the right-hand side of the given differential equation and find $\tilde{y}'_{i+1}$ for column (8).

7th step. We multiply the result of column (8) by the step of integration h and find $h\tilde{y}'_{i+1}$ [column (9)].

8th step. We find Δy_i column (10) for which purpose we find the half-sum of the quantities written in columns (5) and (9).

9th step. We sum up the contents of column (3) with that of column (10) and write the result in column (3) of the next row, i.e. determine $y_{i+1} = y_i + \Delta y_i$.

Then we repeat the whole calculation process beginning with the second step. ▲

The improved Euler-Cauchy method with successive iterative computations. This method is more accurate than the Euler-Cauchy method discussed earlier. The gist of this method is that every value of y_i obtained is subjected to iterative computations. We first find a rough approximation

$$y_{i+1}^{(0)} = y_i + hf(x_i,\ y_i) \tag{8}$$

and then construct an iterative process

$$\bar{y}_{i+1}^{(k)} = y_i + \frac{h}{2}[f(x_i,\ y_i) + f(x_{i+1},\ y_{i+1}^{(k-1)})]. \tag{9}$$

We continue the iterative calculations until two successive approximations $y_{i+1}^{(k)}$ and $y_{i+1}^{(k+1)}$ coincide in the calculation digits in question. Then we assume that $y_{i+1} \cong y_{i+1}^{(k+1)}$. If, after three or four iterations, the digits do not coincide for the chosen value of h, we reduce the calculation step h.

△ **Example 3.** Using the method of iterative computation, find, with the accuracy to within four coincident decimal digits, a solution of the equation $y' = y - x$ for the initial conditions $y(0) = 1$. Obtain the solution of the interval [0, 1.5] take $h = 0.25$.

△ From formula (8) we find that

$$y_1^{(0)} = y_0 + h(y_0 - x_0) = 1.5000 + 0.375 = 1.8750.$$

Next, using the iterative process (9), we find, in succession,

$$\begin{aligned} y_1^{(1)} &= y_0 + \frac{h}{2}\left[(y_0 - x_0) + (y_1^{(0)} - x_1)\right] \\ &= 1.5000 + 0.125\,(1.5000 + 1.875 - 0.25) = 1.89062, \\ y_1^{(2)} &= y_0 + \frac{h}{2}\left[(y_0 - x_0) + (y_1^{(1)} - x_1)\right] \\ &= 1.5000 + 0.125\,(1.5000 + 1.89062 - 0.25) = 1.89258, \\ y_1^{(3)} &= y_0 + \frac{h}{2}\left[(y_0 - x_0) + (y_1^{(2)} - x_1)\right] \\ &= 1.5000 + 0.125\,(1.5000 + 1.89258 - 0.25) = 1.89282, \\ y_1^{(4)} &= y_0 + \frac{h}{2}\left[(y_0 - x_0) + (y_1^{(3)} - x_1)\right] \\ &= 1.5000 + 0.125\,(1.5000 + 1.89282 - 0.25) = 1.89285. \end{aligned}$$

In the last two approximations four digits coincide. Therefore, after rounding off, we can assume $y_1 \cong 1.8929$.

Again using formula (8), for $i = 1$, we find that

$$\begin{aligned} y_2^{(0)} &= y_1 + hf(x_1, y_1) = y_1 + h(y_1 - x_1) \\ &= 1.8929 + 0.25\,(1.8929 - 0.25) = 2.3036. \end{aligned}$$

From formula (9) we find the successive approximations:

$$\begin{aligned} y_2^{(1)} &= 1.8929 + 0.125\,[1.6429 + (2.3036 - 0.50)] = 2.3237, \\ y_2^{(2)} &= 1.8929 + 0.125\,[1.6429 + (2.3237 - 0.50)] = 2.32622, \\ y_2^{(3)} &= 1.8929 + 0.129\,[1.6429 + (2.32622 - 0.50)] = 2.32654, \\ y_2^{(4)} &= 1.8929 + 0.125\,[1.6429 + (2.32654 - 0.50)] = 2.32658. \end{aligned}$$

We can terminate the iterative calculations and assume $y_2 \cong 2.3266$. Using then formulas (8) and (9), we get a solution of the given equation. The results of the calculations are given in Table 9.7. ▲

Table 9.7

i	x_i	y_i	$y_{i+1}^{(0)}$	$y_{i+1}^{(1)}$	$y_{i+1}^{(2)}$	$y_{i+1}^{(3)}$	$y_{i+1}^{(4)}$	y_{i+1}
0	0	1.5000	1.875	1.89062	1.89258	1.89282	1.89285	1.8929
1	0.25	1.8929	2.3036	2.3237	2.32622	2.32654	2.32658	2.3266
2	0.50	2.3266	2.78325	2.80908	2.81231	2.81271	2.81276	2.8128
3	0.75	2.8128	3.3285	3.36171	3.36586	3.3664	3.36645	3.3664
4	1.00	3.3664	3.9580	4.0007	4.00603	4.0067	4.00679	4.0068
5	1.25	4.0068	4.6960	4.7509	4.75776	4.75870	4.75872	4.7587
6	1.50	4.7587						

9.6. The Runge-Kutta Method

The **Runge-Kutta method** is one of the most accurate methods. It has much in common with Euler's method.

Assume that on the interval $[a, b]$ we have to find a numerical solution of the equation

$$y' = f(x, y) \tag{1}$$

with the initial condition

$$y(x_0) = y_0. \tag{2}$$

We divide the interval $[a, b]$ into n equal parts by the points $x_i = x_0 + ih$ $(i = 0, 1, \ldots, n)$, where $h = (b - a)/n$ is the step of integration. As in Euler's method, in the Runge-Kutta method the successive values y_i of the required function y can be found from the formula

$$y_{i+1} = y_i + \Delta y_i. \tag{3}$$

If we expand the function y in a Taylor's series and restrict the calculations to h^4 inclusive, then we can represent the increment of the function Δy in the form

$$\Delta y = y(x+h) - y(x) = hy'(x) + \frac{h^2}{2} y''(x) + \frac{h^3}{6} y'''(x) + \frac{h^4}{24} y^{\mathrm{IV}}(x), \tag{4}$$

where the derivatives $y''(x)$, $y'''(x)$, $y^{\mathrm{IV}}(x)$ can be found from equation (1) by a successive differentiation.

Instead of direct calculations with the use of formula (4), in the Runge-Kutta method we find four numbers:

$$\begin{aligned} k_1 &= hf(x, y), \\ k_2 &= hf\left(x+\frac{h}{2}, y+\frac{k_1}{2}\right), \\ k_3 &= hf\left(x+\frac{h}{2}, y+\frac{k_2}{2}\right), \\ k_4 &= hf(x+h, y+k_3). \end{aligned} \tag{5}$$

We can prove that if we assign the weights 1/6, 1/3 1/3, 1/6 to the numbers k_1, k_2, k_3, k_4 respectively, then with an accuracy to within the fourth powers, the weighted mean of these numbers, i.e.

$$\frac{1}{6}k_1+\frac{1}{3}k_2+\frac{1}{3}k_3+\frac{1}{6}k_4, \tag{6}$$

is equal to the value of Δy calculated from formula (4)

$$\Delta y = \frac{1}{6}(k_1+2k_2+2k_3+k_4). \tag{7}$$

Thus, for every pair of values of x_i and y_i from columns 2 and 3 we find the values of

$$\begin{aligned} k_1^{(i)} &= hf(x_i, y_i), \\ k_2^{(i)} &= hf\left(x_i+\frac{h}{2}, y_i+\frac{k_1^{(i)}}{2}\right), \\ k_3^{(i)} &= hf\left(x_i+\frac{h}{2}, y_i+\frac{k_2^{(i)}}{2}\right), \\ k_4^{(i)} &= hf(x_i+h, y_i+k_3^{(i)}), \end{aligned} \tag{8}$$

from formulas (5) and, using formula (7), we determine

$$\Delta y_i = \frac{1}{6}(k_1^{(i)}+2k_2^{(i)}+2k_3^{(i)}+k_4^{(i)})$$

and then

$$y_{i+1} = y_i + \Delta y_i.$$

The numbers k_1, k_2, k_3, k_4 have a simple geometric meaning. Let the curve M_0CM_1 (Fig. 9.3) be a solution of the differential equation (1) with the initial condition (2). The point C of this curve lies on a straight line which is parallel to the y-axis and bisects the interval

$[x_i, x_{i+1}]$ and B and G are the points of intersection of the tangent drawn to the curve at the point M_0 and the ordinates AC and N_1M_1. Then, with an accuracy to within the factor h (where $h = x_{i+1} - x_i$), the number k_1 is the slope of the tangent at the point M_0 to the integral curve M_0CM_1, i.e. $k_1 = hy_1' = hf(x_i, y_i)$.

The point B has coordinates $x = x_i + \frac{h}{2}$, $y = y_i + \frac{k_1}{2}$ and, consequently, with an accuracy to within the factor h, the number k_2 is the slope of the tangent drawn to the integral curve at the point B (BF is a segment of this tangent).

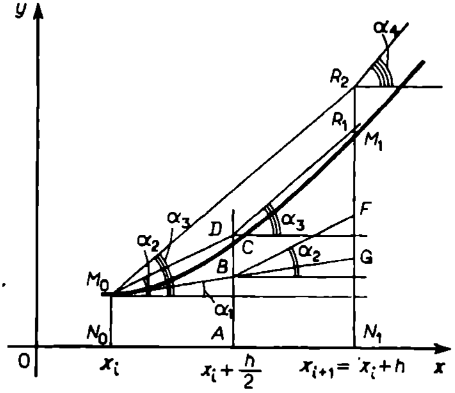


Fig. 9.3

Through the point M_0 we draw a straight line parallel to the segment BF. Then the point D has coordinates $x = x_i + \frac{h}{2}$, $y = y_i + \frac{k_2}{2}$ and k_3 with an accuracy to within the factor h which is the slope of the tangent drawn to the integral curve at the point D (DR_1 is a segment of this tangent). Finally, through the point M_0 we draw a straight line, parallel to DR_1, which cuts the extension of N_1M_1 at the point R_2 $(x_i + h, y_i + k_3)$. Then, with an accuracy to within the factor h, k_4 is the slope of the tangent drawn to the integral curve at the point R_2.

It is convenient to use the scheme shown in Table 9.8

to make calculations by the Runge-Kutta method. This table is filled in as follows.

1st step. In columns (2) and (3) of the running row

Table 9.8

i	x	y	$y' = f(x, y)$	$k = hf(x, y)$	Δy
(1)	(2)	(3)	(4)	(5)	(6)
0	x_0	y_0	$f(x_0, y_0)$	$k_1^{(0)}$	$k_1^{(0)}$
	$x_0 + \frac{h}{2}$	$y_0 + \frac{k_1^{(0)}}{2}$	$f\left(x_0 + \frac{h}{2}, y_0 + \frac{k_1^{(0)}}{2}\right)$	$k_2^{(0)}$	$k_2^{(0)}$
	$x_0 + \frac{h}{2}$	$y_0 + \frac{k_2^{(0)}}{2}$	$f\left(x_0 + \frac{h}{2}, y_0 + \frac{k_2^{(0)}}{2}\right)$	$k_3^{(0)}$	$2k_3^{(0)}$
	$x_0 + h$	$y_0 + k_3^{(0)}$	$f(x_0 + h, y_0 + k_3^{(0)})$	$k_4^{(0)}$	$k_4^{(0)}$
					$\frac{1}{6}\sum = \Delta y_0$
1	x_1	$y_1 = y_0 + \Delta y_0$	$f(x_1, y_1)$	$k_1^{(1)}$	$k_1^{(1)}$
	$x_1 + \frac{h}{2}$	$y_1 + \frac{k_1^{(1)}}{2}$	$f\left(x_1 + \frac{h}{2}, y_1 + \frac{k_1^{(1)}}{2}\right)$	$k_2^{(1)}$	$2k_2^{(1)}$
	$x_1 + \frac{h}{2}$	$y_1 + \frac{k_2^{(1)}}{2}$	$f\left(x_1 + \frac{h}{2}, y_1 + \frac{k_2^{(1)}}{2}\right)$	$k_3^{(1)}$	$2k_3^{(1)}$
	$x_1 + h$	$y_1 + k_3^{(1)}$	$f(x_1 + h, y_1 + k_3^{(1)})$	$k_4^{(1)}$	$k_4^{(1)}$
					$\frac{1}{6}\sum = \Delta y_1$
2	x_2	$y_2 = y_1 + \Delta y_1$			

we write the values of x and y we need. (If it is the first row, then we write the initial data, x_0 and y_0.)

2nd step. We substitute the values of x and y of the running row into the right-hand side of the differential equation (1), find $f(x, y)$ and write the result in column (4) of the same row.

3rd step. We multiply the value of $f(x, y)$ of column (4) we have obtained by the step of integration h, calculate $k = hf(x, y)$ and write the result in column (5) of the same row.

4th step. We multiply the values of k obtained by the appropriate coefficient (by 1 if the value is k_1 or k_4 or by 2 if it is k_2 or k_3) and write the result in column (6) of the running row.

We repeat steps 1, 2, 3 and 4 to find every k in the ith solution.

We sum up the results of the sixth column, divide by 6 and find $\Delta y_i = \frac{1}{6}\Sigma$ and $y_{i+1} = y_i + \Delta y_i$.

Then we repeat all the calculations beginning with the first step until we cover the whole interval $[a, b]$.

The Runge-Kutta method has the order of accuracy h^4 throughout the interval $[a, b]$. It is very difficult to estimate the accuracy of this method. We can get a rough estimate of the error by means of the "duplication check" from the formula

$$|y_i^* - y(x_i)| \cong \frac{y_i^* - y_i}{15} \tag{9}$$

where $y(x_i)$ is the value of the exact solution of equation (1) at the point x_i and y_i^* and y_i are approximate values obtained with the steps $h/2$ and h.

If ε is the preassigned accuracy of the solution, then the number n (the number of partitions) for determining the step of integration $h = (b - a)/n$ is chosen such that

$$h^4 < \varepsilon. \tag{10}$$

However, we can change the step of calculations when we pass from one point to another.

We can estimate the correctness of the choice of the step h using the relation

$$q = \left| \frac{k_2^{(i)} - k_3^{(i)}}{k_1^{(i)} - k_2^{(i)}} \right| \tag{11}$$

where q must be equal to several hundredths, otherwise we must reduce the stepsize h.

Example 1. Given a differential equation $y' = y - x$ satisfying the initial condition $y(0) = 1.5$, find, with an accuracy of 0.01 a solution of this equation for $x = 1.5$. Use the Runge-Kutta method and make calculations with two extra digits.

△ We choose the initial step of calculations h from the condition $h^4 < 0.01$. Then $h < 0.3$. To make the calculations more convenient, we assume that $h = 0.25$. We divide the whole integration interval $[0, 1.5]$ into six equal parts by the points $x_0 = 0$, $x_1 = 0.25$, $x_2 = 0.50$, $x_3 = 0.75$, $x_4 = 1.00$, $x_5 = 1.25$, $x_6 = 1.50$. From the initial conditions we have $x_0 = 0$, $y_0 = 1.5$. We seek the first approximation $y_1 + \Delta y_0$, where

$$\Delta y_0 = \frac{1}{6}\left(k_1^{(0)} + 2k_2^{(0)} + 2k_3^{(0)} + k_4^{(0)}\right).$$

Using formulas (8), we obtain

$$k_1^{(0)} = (y_0 - x_0)\,h = 1.5000 \cdot 0.25 = 0.3750,$$

$$k_2^{(0)} = \left[\left(y_0 + \frac{k_1^{(0)}}{2}\right) - \left(x_0 + \frac{h}{2}\right)\right] h$$

$$= [(1.5000 + 0.1875) - 0.125] \cdot 0.25 = 0.3906.$$

$$k_3^{(0)} = \left[\left(y_0 + \frac{k_2^{(0)}}{2}\right) - \left(x_0 + \frac{h}{2}\right)\right] h$$

$$= [(1.5000 + 0.1953) - 0.125] \cdot 0.25 = 0.3926.$$

$$k_4^{(0)} = [(y_0 + k_3^{(0)}) - (x_0 + h)]\,h$$

$$= [(1.5000 + 0.3926) - 0.25] \cdot 0.25 = 0.4106.$$

Consequently,

$$\Delta y_0 = \frac{1}{6}(0.3750 + 2 \cdot 0.3906 + 2 \cdot 0.3926 + 0.4106) = 0.3920,$$

$$y_1 = 1.5000 + 0.3920 = 1.8920.$$

One can see from Table 9.9 how to go on with the solution of the equation. Thus the final result is $y(1.5) = 4.74$. ▲

The Runge-Kutta method can be employed in the solution of systems of differential equations.

Consider a system of first-order differential equations

$$\begin{cases} y' = f(x, y, z), \\ z' = g(x, y, z) \end{cases} \tag{12}$$

with the initial conditions

$$x = x_0,\ y(x_0) = y_0,\ z(x_0) = z_0. \tag{13}$$

In this case we determine the numbers Δy_i and Δz_i:

$$\begin{aligned} \Delta y_i &= \frac{1}{6}(k_1^{(i)} + 2k_2^{(i)} + 2k_3^{(i)} + k_4^{(i)}, \\ \Delta z_i &= \frac{1}{6}(l_1^{(i)} + 2l_2^{(i)} + 2l_3^{(i)} + l_4^{(i)}, \end{aligned} \tag{14}$$

Table 9.9

i	x	y	$y' = f(x, y)$	$k = hf(x, y)$	Δy
(1)	(2)	(3)	(4)	(5)	(6)
0	0	1.5000	1.5000	0.3750	0.3750
	0.125	1.6875	1.5625	0.3906	0.7812
	0.125	1.6953	1.5703	0.3926	0.7852
	0.25	1.8926	1.6426	0.4106	0.4106
					0.3920
1	0.25	1.8920	1.6420	0.4105	0.4105
	0.375	2.0973	1.7223	0.4306	0.8612
	0.375	2.1073	1.7323	0.4331	0.8662
	0.50	2.3251	1.8251	0.4562	0.4562
					0.4323
2	0.50	2.2343	1.8243	0.4561	0.4561
	0.625	2.5523	1.9273	0.4818	0.963
	0.625	2.5652	1.9402	0.4850	0.9700
	0.75	2.8093	2.0593	0.5148	0.5148
					0.4841
3	0.75	2.8084	2.0584	0.5146	0.5146
	0.875	3.0657	2.1907	0.5477	1.0954
	0.875	3.0823	2.2073	0.5518	1.1036
	1.00	3.3602	2.3602	0.5900	0.5900
					0.5506
4	1.00	3.3590	2.3590	0.5898	0.5898
	1.125	3.6539	2.5289	0.6322	1.2641
	1.125	3.6751	2.5501	0.6375	1.2750
	1.25	3.9965	2.7465	0.6686	0.6866
					0.6360

Table 9.9 (continued)

i	x	y	$y' = f(x, y)$	$k = hf(x, y)$	Δy
(1)	(2)	(3)	(4)	(5)	(6)
5	1.25	3.9950	2.7450	0.6862	0.6862
	1.375	4.3381	2.9631	0.7408	1.4816
	1.375	4.3654	2.9904	0.7476	1.4952
	1.50	4.7426	3.2426	0.8106	0.8106
					0.7456
6	1.50	4.7406			

where

$$k_1^{(i)} = hf(x_i, y_i, z_i),$$
$$l_1^{(i)} = hg(x_i, y_i, z_i),$$
$$k_2^{(i)} = hf\left(x_i + \frac{h}{2},\ y_i + \frac{k_1^{(i)}}{2},\ z_i + \frac{l_1^{(i)}}{2}\right),$$
$$l_2^{(i)} = hg\left(x_i + \frac{h}{2},\ y_i + \frac{k_1^{(i)}}{2},\ z_i + \frac{l_1^{(i)}}{2}\right),$$
$$k_3^{(i)} = hf\left(x_i + \frac{h}{2},\ y_i + \frac{k_2^{(i)}}{2},\ z_i + \frac{l_2^{(i)}}{2}\right),$$
$$l_3^{(i)} = hg\left(x_i + \frac{h}{2},\ y_i + \frac{k_2^{(i)}}{2},\ z_i + \frac{l_2^{(i)}}{2}\right),$$
$$k_4^{(i)} = hf(x_i + h,\ y_i + k_3^{(i)},\ z_i + l_3^{(i)}).$$
$$l_4^{(i)} = hg(x_i + h,\ y_i + k_3^{(i)},\ z_i + l_3^{(i)}).$$

Then we get a solution of the system

$$y_{i+1} = y_i + \Delta y_i,\quad z_{i+1} = z_i + \Delta z_i.$$

Example 2. Given a system of differential equations

$$\begin{cases} y' = \dfrac{2y - x}{z} \\ z' = \dfrac{2y}{z + x} \end{cases}$$

with the initial conditions $x_0 = 0.5$, $y_0 = 1$, $z_0 = 1$, find a solution of the system for $x = 0.6$. Make calculations with five decimal digits.

△ We choose a step $h = 0.1$ and find the numbers k_1, l_1, k_2, l_2, k_3, l_3, k_4, l_4:

$$k_1 = h\cdot\frac{2y_0 - x_0}{z_0} = 0.1\cdot\frac{2-0.5}{1} = 0.15000,$$

$$l_1 = h\cdot\frac{2y_0}{z_0+x_0} = 0.1\cdot\frac{2}{1.5} = 0.13333,$$

$$k_2 = h\left[\frac{2\left(y_0+\frac{k_1}{2}\right)-\left(x_0+\frac{h}{2}\right)}{z_0+\frac{l_1}{2}}\right] = 0.1\cdot\frac{2\cdot 1.075-0.55}{1.06667} = 0.14100,$$

$$l_2 = h\left[\frac{2\left(y_0+\frac{k_1}{2}\right)}{\left(z_0+\frac{l_1}{2}\right)+\left(x_0+\frac{h}{2}\right)}\right] = 0.1\cdot\frac{2\cdot 1.075}{1.06667+0.55} = 0.13299,$$

$$k_3 = h\left[\frac{2\left(y_0+\frac{k_2}{2}\right)-\left(x_0+\frac{h}{2}\right)}{z_0+\frac{l_2}{2}}\right]$$

$$= 0.1\cdot\frac{2\cdot 1.07050-0.55}{\cdot 1.06650} = 0.14918,$$

$$l_3 = h\left[\frac{2\left(y_0+\frac{k_2}{2}\right)}{\left(z_0+\frac{l_2}{2}\right)+\left(x_0+\frac{h}{2}\right)}\right] = 0.1\cdot\frac{2\cdot 1.07050}{1.06650+0.55} = 0.13245,$$

$$k_4 = h\left[\frac{2(y_0+k_3)-(x_0+h)}{z_0+l_3}\right] = 0.1\cdot\frac{2\cdot 1.14918-0.16}{1.13245} = 0.14998,$$

$$l_4 = h\left[\frac{2(y_0+k_3)}{(z_0+l_3)+(z_0+h)}\right] = 0.1\cdot\frac{2\cdot 1.14918}{1.13245+0.6} = 0.13266,$$

Consequently,

$$\Delta y_0 = \frac{1}{6}(0.15+2\cdot 0.14100+2\cdot 0.14918+0.14998) = 0.14672,$$

$$\Delta z_0 = \frac{1}{6}(0.13333+2\cdot 0.13299+2\cdot 0.13245+0.13266) = 0.13281$$

and we get the value of the required functions at the point $x = 0.6$: $y_1 = 1 + 0.14672 = 1.14672$, $z_1 = 1 + 0.13281 = 1.13281$. ▲

9.7. Adams' Extrapolation Method

When using the Runge-Kutta method to solve a differential equation, it is necessary to carry out numerous calculations to find every y_i. In the case when the right-hand side of an equation includes a complicated analytic expression, the solution of the equation by means of the Runge-Kutta method is very laborious. Therefore, in practical computations, **Adams' method** is usually used which does not require repeated calculations of the right-hand side of the equation.

Consider a differential equation

$$y' = f(x, y) \tag{1}$$

with the initial condition

$$x = x_0, \ y(x_0) = y_0. \tag{2}$$

We have to find a solution of this equation on the interval $[a, b]$.

We divide the interval $[a, b]$ into n equal parts by the points $x_i = x_0 + ih$ $(i = 0, 1, 2, \ldots, n)$. We choose a subinterval $[x_i, x_{i+1}]$ and integrate the differential equation (1). Then we get

$$y_{i+1} = y_i + \int_{x_i}^{x_{i+1}} y' \, dx, \tag{3}$$

or

$$\Delta y_i = \int_{x_i}^{x_{i+1}} y' \, dx.$$

To find the derivative, we use Newton's second interpolation formula (restricting our computations to third-order differences):

$$y' = y'_i + t\Delta y'_{i-1} + \frac{t(t+1)}{2!} \Delta^2 y'_{i-2} + \frac{t(t+1)(t+2)}{3!} \Delta^3 y'_{i-3}, \tag{4}$$

where $t = (x - x_i)/h$, or

$$y' = y'_i + t\Delta y'_{i-1} + \frac{t^2 + t}{2} \Delta^2 y'_{i-2} + \frac{t^3 + 3t^2 + 2}{6} \Delta^3 y'_{i-3}. \tag{4'}$$

Substituting the expression for y' from formula (4') into relation (3) and taking into account that $dx = h\,dt$, we have

$$\Delta y_i = h \int_0^1 \left(y'_i + t\Delta y'_{i-1} + \frac{t^2+t}{2} \Delta^2 y'_{i-2} \right.$$

$$\left. + \frac{t^3+3t^2+2t}{6} \Delta^3 y'_{i-3} \right) dt$$

$$= hy'_i + \frac{1}{2} \Delta (hy'_{i-1}) + \frac{5}{12} \Delta^2 (hy'_{i-2}) + \frac{3}{8} \Delta^3 (hy'_{i-3}). \tag{5}$$

In what follows, we designate

$$q_i = y'_i h = f(x_i, y_i) \cdot h \ (i = 0, 1, 2, \ldots, n).$$

Then, for any difference, we have $\Delta^m q_i = \Delta^m (y'_i h)$ and

$$\Delta y_i = q_i + \frac{1}{2} \Delta q_{i-1} + \frac{5}{12} \Delta^2 q_{i-2} + \frac{3}{8} \Delta^3 q_{i-3}. \tag{6}$$

From the formula $y_{i+1} = y_i + \Delta y_i$ we get a solution of the equation. Formula (6) is known as *Adams' extrapolation formula.*

To begin the computation process, we need four initial values y_0, y_1, y_2, y_3 constituting the so-called ***initial interval*** which can be found from the initial condition (2) with the use of one of the familiar methods. The initial interval of the solution is usually obtained by means of the Runge-Kutta method.

Knowing y_0, y_1, y_2, y_3, we can determine

$$q_0 = hy'_0 = hf(x_0, y_0); \ q_1 = hy'_1 = hf(x_1, y_1),$$

$$q_2 = hy'_2 = hf(x_2, y_2); \ q_3 = hy'_3 = hf(x_3, y_3). \tag{7}$$

Then we compile a table of differences involving the quantity q (Table 9.10).

Adams' method consists in extending the diagonal table of differences with the use of formula (6). By means of the numbers q_3, Δq_2, $\Delta^2 q_1$, $\Delta^3 q_0$, which cross the table along the diagonal, we find from formula (6), setting $n = 3$ in it (the last known value of y is y_3), that

$$\Delta y_3 = q_3 + \frac{1}{2} \Delta q_2 + \frac{5}{12} \Delta^2 q_1 + \frac{3}{8} \Delta^3 q_0.$$

Table 9.10

i	x_i	y_i	Δy_i	$y'_i = f(x_i, y_i)$	$q = hy'_i$	Δq_i	$\Delta^2 q_i$	$\Delta^3 q_i$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
0	x_0	y_0		$f(x_0, y_0)$	q_0	Δq_0	$\Delta^2 q_0$	$\Delta^3 q_0$
1	x_1	y_1		$f(x_1, y_1)$	q_1	Δq_1	$\Delta^2 q_1$	
2	x_2	y_2		$f(x_2, y_2)$	q_2	Δq_2		
3	x_3	y_3	Δy_3	$f(x_3, y_3)$	q_3			
4	x_4	y_4						
5	x_5							
6	x_6							

We tabulate the value of Δy_3 and find $y_4 = y_3 + \Delta y_3$. Then, using x_4 and the value of y_4 we have found, we determine $f(x_4, y_4)$, q_4, Δq_3, $\Delta^2 q_2$, $\Delta^3 q_1$, i.e. we get a new diagonal. From these data we obtain

$$\Delta y_4 = q_3 + \frac{1}{2}\Delta q_3 + \frac{5}{12}\Delta^2 q_2 + \frac{3}{8}\Delta^3 q_1, \quad y_5 = y_4 + \Delta y_4.$$

We thus continue with the table of solution calculating the right-hand side of the differential equation (1) once at each stage.

We can also use Runge's principle to make a rough estimate of the error. It consists in the following:

(1) we find a solution of the differential equation for the step h,

(2) next we double the value of the step h and find the solution for the step $H = 2h$,

(3) and then we calculate the error of the method from the formula

$$\varepsilon = \frac{|\tilde{y}_h - \tilde{y}_{2h}|}{2^m - 1}, \qquad (8)$$

where $\bar{y}_n$ is the value of the approximate calculation for the double step $H = 2h$ and $\tilde{y}_{2n}$ is the value of the approximate calculation for the step h.

Remark. When calculating with the step h, we assume that we have an error proportional to h^{m+1} at every step and with the step $2h$ we have an error proportional to $(2h)^{m+1}$ if the order of accuracy of the method is defined and equal to h^m.

Note that in Adams' extrapolation formula (6) the third finite differences $\Delta^2 q$ are assumed to be constant. Therefore we can find the value h of the initial step of calculations from the inequality $h^4 < \varepsilon$, where ε is the preassigned accuracy of the solution.

In practical calculations we usually watch the course of the third finite differences choosing h such that the adjoining differences $\Delta^3 q_i$ and $\Delta^3 q_{i+1}$ should differ from each other by not more than one or two unities of the specified decimal place (not counting the extra digits).

Example 1. Use Adams' method to calculate, with an accuracy of 0.01, the value of the solution of the differential equation $y' = y - x$ for the initial $x_0 = 0$ and $y_0 = 1.5$ when $x = 1.5$. Carry out all calculations with two extra digits.

△ As before, we choose h from the relation $h^4 < 0.01$, i.e. $h = 0.25$. We take the initial interval y_0, y_1, y_2, y_3 from the solution of Example 1 in 9.6. To solve this equation, we compile two tables, the main Table 9.11 and the auxiliary Table 9.12. Their purpose is clear from the tables themselves.

The final result is y (1.5) = 4.74. ▲

Table 9.11

i	x_i	y_i	Δy_i	$y'_i = f(x_i, y_i)$	$q_i = hy'_i$	Δq_i	$\Delta^2 q_i$	$\Delta^3 q_i$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
0	0	1.5000		1.5000	0.3750	0.0355	0.0101	0.0028
1	0.25	1.8920		1.6420	0.4105	0.0456	0.0129	0.0037
2	0.50	2.3243		1.8243	0.4561	0.0585	0.0166	0.0047
3	0.75	2.8084	0.5504	2.0584	0.5146	0.0751	0.0213	
4	1.00	3.3588	0.6356	2.3588	0.5897	0.0964		
5	1.25	3.9944	0.7450	2.7444	0.6861			
6	1.50	4.7394						

Table 9.12

i	q_i	$\frac{1}{2}\Delta q_{i-1}$	$\frac{5}{12}\Delta^2 q_{i-2}$	$\frac{3}{8}\Delta^3 q_{i-3}$	Δy_i
3	0.5146	0.0293	0.0054	0.0011	0.5504
4	0.5897	0.0376	0.0069	0.0014	0.6356
5	0.6861	0.0482	0.0089	0.0018	0.7450

Adams' method can also be used to solve systems of differential equations and nth-order differential equations.

Assume that we have a system of two equations

$$\begin{cases} y' = f_1(x, y, z), \\ z' = f_2(x, y, z). \end{cases} \tag{9}$$

Then Adams' extrapolation formulas for this system have the form

$$\begin{aligned} \Delta y_i &= p_i + \frac{1}{2}\Delta p_{i-1} + \frac{5}{12}\Delta^2 p_{i-2} + \frac{3}{8}\Delta^3 p_{i-3}, \\ \Delta z_i &= g_i + \frac{1}{2}\Delta g_{i-1} + \frac{5}{12}\Delta^2 g_{i-2} + \frac{3}{8}\Delta^3 g_{i-3}, \end{aligned} \tag{10}$$

where

$$p_i = hy'_i = hf_1(x_i, y_i, z_i),$$
$$g_i = hz'_i = hf_2(x_i, y_i, z_i)$$

and

$$y_{i+1} = y_i + \Delta y_i, \quad z_{i+1} = z_i + \Delta z_i.$$

Example 2. Using Adams' method, find a numerical solution of the system of differential equations

$$\begin{cases} y' = (z - y)\, x, \\ z' = (z + y)\, x \end{cases}$$

for the initial conditions $y(0) = 1.000$, $z(0) = 1.000$ on the interval $[0, 0.6]$, the step is $h = 0.1$.

△ We take the initial interval of solution from Table 9.3 (earlier we used Euler's method to solve this system). We shall seek the values of the functions $y(x)$ and $z(x)$ for $x_4 = 0.4$, $x_5 = 0.5$ and $x_6 = 0.6$ using formulas (10) and designating $f_1(x, y, z) = y' = (z - y)\,x$ and $f_2(x, y, z) = z' = (z + y)\,x$. The calculations are given in Tables 9.13, 9.14, and 9.15 (Tables 9.14 and 9.15 are auxiliary).

Table 9.13

i	x_i	y_i	Δy_i	p_i	Δp_i	$\Delta^2 p_i$	$\Delta^3 p_i$
0	0	1.0000		0.0000	0.0000	0.0004	0.0006
1	0.1	1.0000		0.0000	0.0004	0.0010	0.0010
2	0.2	1.0000		0.0004	0.0014	0.0020	0.0004
3	0.3	1.0004	0.0032	0.0018	0.0034	0.0024	
4	0.4	1.0036	0.0081	0.0052	0.0058		
5	0.5	1.0117	0.0150	0.0110			
6	0.6	1.0267					

i	x_i	z_i	Δz_i	g_i	Δg_i	$\Delta^2 g_i$	$\Delta^3 g_i$
0	0	1.0000		0.0000	0.0200	0.0004	0.0006
1	0.1	1.0000		0.0200	0.0204	0.0010	0.0013
2	0.2	1.0200		0.0404	0.0214	0.0023	0.0007
3	0.3	1.0604	0.0732	0.0618	0.0237	0.0030	
4	0.4	1.1336	0.0984	0.0855	0.0267		
5	0.5	1.2323	0.1271	0.1122			
6	0.6	1.3594					

Table 9.14

i	x_i	y_i	z_i	y'_i	p_i	z'_i	g_i
0	0.0	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000
1	0.1	1.0000	1.0000	0.0000	0.0000	0.2000	0.0200
2	0.2	1.0000	1.0200	0.0040	0.0004	0.4040	0.0404
3	0.3	1.0004	1.0604	0.0180	0.0018	0.6182	0.0618
4	0.4	1.0036	1.1336	0.0520	0.0052	0.8549	0.0855
5	0.5	1.0117	1.2323	0.1103	0.0110	0.1220	0.1122

Table 9.15

i	p_i	$\frac{1}{2}\Delta p_{i-1}$	$\frac{5}{12}\Delta^2 p_{i-2}$	$\frac{3}{8}\Delta^3 p_{i-3}$	Δy_i
3	0.0018	0.0007	0.000425	0.000225	0.0032
4	0.0052	0.0017	0.000850	0.000375	0.0081
5	0.0110	0.0029	0.0010	0.00015	0.0150

Table 9.15 (continued)

i	g_i	$\frac{1}{2}\Delta g_{i-1}$	$\frac{5}{12}\Delta^2 g_{i-2}$	$\frac{3}{8}\Delta^3 g_{i-3}$	Δz_i
3	0.0618	0.0107	0.000425	0.000225	0.0732
4	0.0855	0.0118	0.000958	0.000488	0.0987
5	0.1122	0.0134	0.00125	0.00026	1.1271

Table 9.14 serves for determining the right-hand sides of this system and for finding p_i and g_i, and Table 9.15, for determining Δy_i and Δz_i from the differences of the values of p and g obtained in Table 9.13. ▲

9.8. Milne's Method

As the Runge-Kutta method, Milne's method is of high accuracy.

Assume that we have to find, on the interval $[a, b]$, a numerical solution of the differential equation

$$y' = f(x, y) \tag{1}$$

with the initial condition

$$y(x_0) = y_0. \tag{2}$$

We divide the interval $[a, b]$ into n equal parts by the points $x_i = x_0 + ih$ $(i = 0, 1, \ldots, n)$, where $h = (b - a)/n$ is the step of integration.

Using the initial data, we employ some technique to find successive values $y_1 = y(x_1)$, $y_2 = y(x_2)$, $y_3 = y(x_3)$ of the required function $y(x)$. Thus we find y'_i $(i = 0, 1, 2, 3)$.

We can find the approximation $\bar{y}_i$ and $\bar{\bar{y}}_i$ for the successive values y_i $(i = 4, 5, \ldots, n)$ from *Milne's formulas*:

$$\bar{y}_i = y_{i-4} + \frac{4h}{3}(2y'_{i-3} - y'_{i-2} + 2y'_{i-1}), \tag{3}$$

$$\bar{\bar{y}}_i = y_{i-2} + \frac{h}{3}(\bar{y}'_i + 4y'_{i-1} + y'_{i-2}), \tag{4}$$

where $\bar{y}'_i = f(x_i, \bar{y}_i)$.

We can show that the absolute error of the value $\bar{\bar{y}}_i$ is approximately equal to

$$\varepsilon_i = \frac{1}{29}|\bar{\bar{y}}_i - \bar{y}_i|. \tag{5}$$

Therefore, if $\varepsilon_i \leqslant \varepsilon$, where ε is the preassigned limiting error of the solution, then we can set $y_i \cong \bar{y}_i$ and $y'_i = f(x_i, \bar{\bar{y}}_i)$. This occurs when $\bar{y}_i$ and $\bar{\bar{y}}_i$ coincide at the decimal places of interest to us. If condition (5) is fulfilled, we pass to the calculation of the next value y_{i+1}, repeating the process. Otherwise, beginning with a certain place, we reduce the stepsize h and recalculate the corresponding initial interval. As in the Runge-Kutta method, the value of the initial step can be found from the inequality $h^4 < \varepsilon$.

To derive Milne's formulas (3) and (4), we use the Newton's first interpolation formula for the derivative y' at a chosen point x_k and restrict our calculations to the third-order differences. This is equivalent to the approximation of the integral $y = y(x)$ of the differential equation (1) by a four-degree polynomial. We have

$$y' = y'_k + q\Delta y'_k + \frac{q(q-1)}{2!}\Delta^2 y'_k + \frac{q(q-1)(q-2)}{3!}\Delta^3 y'_k. \qquad (6)$$

Removing the brackets, we have

$$y' = y'_k + q\Delta y'_k + \frac{1}{2}(q^2-q)\Delta^2 y'_k + \frac{1}{6}(q^3-3q^2+2q)\Delta^3 y'_k, \qquad (7)$$

where $q = (x-x_k)/h$.

Setting $k = i-4$ in relation (7), we obtain

$$\begin{aligned} y' = y'_{i-4} + q\Delta y'_{i-4} + \frac{1}{2}(q^2-q)\Delta^2 y'_{i-4} \\ + \frac{1}{6}(q^3-3q^2+2q)\Delta^3 y'_{i-4}. \end{aligned} \qquad (8)$$

We integrate relation (8) with respect to x from x_{i-4} to x_i:

$$\begin{aligned} \int_{x_{i-4}}^{x_i} y'\,dx = \int_{x_{i-4}}^{x_i} \Big[y'_{i-4} + q\Delta y'_{i-4} + \frac{q^2-q}{2}\Delta^2 y'_{i-4} \\ + \frac{q^3-3q^2+2q}{6}\Delta^3 y'_{i-4} \Big] dx. \end{aligned}$$

Taking into account that $q = (x - x_{i-4})/h$ and $dx = h\,dq$, we have

$$\begin{aligned} y_i - y_{i-4} = h\Big\{ y'_{i-4}\int_0^4 dq + \Delta y'_{i-4}\int_0^4 q\,dq + \Delta^2 y'_{i-4}\int_0^4 \frac{q^2-q}{2}\,dq \\ + \Delta^3 y'_{i-4}\int_0^4 \frac{q^3-3q^2+2q}{6}\,dq \Big\} \\ = h\Big(4y'_{i-4} + 8\Delta y'_{i-4} + \frac{20}{3}\Delta^2 y'_{i+4} + \frac{8}{3}\Delta^3 y'_{i-4}\Big). \end{aligned} \qquad (9)$$

Since

$$\Delta y'_{i-4} = y'_{i-3} - y'_{i-4},$$

$$\Delta^2 y'_{i-4} = y'_{i-2} - 2y'_{i-3} + y'_{i-4},$$

$$\Delta^3 y'_{i-4} = y'_{i-1} - 3y'_{i-2} + 3y'_{i-3} - y'_{i-4},$$

we can substitute these expressions into relation (9) and get Milne's first formula (3):

$$\bar{y}_i = y_{i-4} + \frac{4h}{3}(2y'_{i-3} - y'_{i-2} + 2y'_{i-1}).$$

To derive Milne's second formula (4), we set $k = i - 2$ in relation (7):

$$y' = y'_{i-2} + q\Delta y'_{i-2} + \frac{1}{2}(q^2 - q)\Delta^2 y'_{i-2}$$
$$+ \frac{1}{6}(q^3 - 3q^2 + 2q)\Delta^3 y'_{i-2}. \tag{10}$$

where $q = (x - x_{i-2})/h$ and $dx = h\,dq$.

We integrate relation (10) with respect to x from x_{i-2} to x_i:

$$\int_{x_{i-2}}^{x_i} y'\,dx = h\int_0^2 \left[y'_{i-2} + q\Delta y'_{i-2} + \frac{1}{2}(q^2 - q)\Delta^2 y'_{i-2}\right.$$
$$\left. + \frac{1}{6}(q^3 - 3q^2 + 2q)\Delta^3 y'_{i-2}\right] dq.$$

From this we have

$$y_i - y_{i-2} = h\left(2y'_{i-2} + 2\Delta y'_{i-2} + \frac{1}{3}\Delta^2 y'_{i-2}\right). \tag{11}$$

Taking into account that $\Delta y'_{i-2} = y'_{i-1} - y'_{i-2}$ and $\Delta^2 y'_{i-2} = y'_i - 2y'_{i-1} + y'_{i-2}$, we get Milne's second formula (4):

$$\bar{\bar{y}}_i = y_{i-2} + \frac{h}{3}(y'_i + 4y'_{i-1} + y'_{i-2}).$$

Example 1. Given a differential equation $y' = y - x$ satisfying the initial condition $x_0 = 0$, $y'(x_0) = 1.5$, calculate, with an accuracy of 0.01, the value of the solution of this equation for $x = 1.5$. Carry out calculations by the combined Runge-Kutta and Milne's method with two extra digits.

△ We choose the initial step of calculations. From the condition $h^4 < 0.01$ we get $h = 0.25$. Then we can divide the whole integration interval into six equal parts by the points $x_0 = 0$, $x_1 = 0.25$, $x_2 = 0.50$, $x_3 = 0.75$, $x_4 = 1.00$, $x_5 = 1.25$, $x_6 = 1.50$. We take the initial interval y_0, y_1, y_2, y_3 from the solution of Example 1 in 9.6. To solve this equation, we compile Table 9.16.

Table 9.16

i	x_i	y_i	$y'_i = f(x_i, y_i) = y_i - x_i$	$\bar{y}_i$	$\bar{y}'_i = f(x_i, \bar{y}_i) = \bar{y}_i - x_i$	$\bar{\bar{y}}_i$	ε_i	y_i	$y'_i = f(x_i, y_i) = y_i - x_i$	The recalculation of step via the result of formula (5)
0	0	1.5000	1.5000							
1	0.25	1.8920	1.6420							
2	0.50	2.3243	1.8243							
3	0.75	2.8084	2.0584							
4	1.00			3.3588	2.3588	3.3590	$7\cdot10^{-5}$	3.3590	2.3590	Not necessary
5	1.25			3.9947	2.7447	3.9950	10^{-6}	3.9950	2.7450	» »
6	1.50			4.7402	3.2402	4.7406	$1.4\cdot10^{-5}$	4.7406		» »

Thus we got an answer: $y(1.5) = 4.74$. ▲

When solving a system of differential equations

$$\begin{cases} y' = f(x, y, z), \\ z' = \varphi(x, y, z) \end{cases}$$

with the initial conditions $y(x_0) = y_0$, $z(x_0) = z_0$, Milne's formulas are written separately for the functions $y(x)$ and $z(x)$. The order of calculations remains the same.

Example 2. Given a system of equations

$$\begin{cases} y' = \cos(y+1,\ 1z)+1, \\ z' = \dfrac{1}{x+2.1y^2}+x+1 \end{cases}$$

with the initial conditions $y(0) = 3.14159$, $z(0) = 0$. Use Milne's method to calculate, with an accuracy of $\varepsilon = 0.0001$, the solution

Table 9.17

(1)	(2)	(3)	(4)	(5)	(6)
i	x_i	$\bar{Y}_i = (y_i, z_i)$		$\bar{Y}'_i$	
		$\bar{y}_i$	$\bar{z}_i$	$\bar{y}'_i$	$\bar{z}'_i$
0	0.0				
1	0.1				
2	0.2				
3	0.3				
4	0.4	3.16057	0.49906	0.15698	1.44678
5	0.5	3.18164	0.64870	0.27079	1.54596

(7)	(8)	(9)	(10)	(11)	(12)
$\bar{\bar{Y}}_i = Y_i$		$\bar{\bar{Y}}'_i = Y'_i$		ε_i	
$\bar{\bar{y}}_i$	$\bar{\bar{z}}_i$	$\bar{\bar{y}}'_i = y'_i$	$\bar{\bar{z}}'_i = z'_i$	$\varepsilon_{i,1}$	$\varepsilon_{i,2}$
3.14159	0.00000	0.00000	1.04824		
3.14184	0.10981	0.00732	1.14801		
3.14364	0.22960	0.03224	1.24773		
3.14903	0.35934	0.08001	1.34734		
3.16062	0.49905	0.15701	1.44678	$1.7\cdot 10^{-6}$	0
3.18166	0.64869			7.10^{-7}	$4\cdot 10^{-7}$

of this system on the interval [0, 0.5] taking a step $h = 0.1$ and assuming the values of the functions $y(x)$ and $z(x)$ to be known for $x_1 = 0.1$, $x_2 = 0.2$, $x_3 = 0.3$. These values are $y(0.1) = 3.14184$, $y(0.2) = 3.14364$, $y(0.3) = 3.14903$, $z(0.1) = 0.10981$, $z(0.2) = 0.22960$, $z(0.3) = 0.35934$.

△ We shall seek the values of the functions $y(0.4)$, $y(0.5)$ and $z(0.4)$, $z(0.5)$ from formulas (3)-(5). We carry out the calculations with the aid of two tables: the main Table 9.17 and the auxiliary Table 9.18.

We write the initial interval of the solution in columns (7) and (8) of Table 9.17 and find the values of y'_i and z'_i [columns (9)

Table 9.18

	4		5	
	y_i	z_i	y_i	z_i
$2Y'_{i-3}$	0.01464	2.29602	0.06448	2.49546
$-Y'_{i-2}$	−0.30224	−1.24773	−0.08001	−1.34734
$2Y'_{i-1}$	0.16002	2.69468	0.31402	2.89356
$\sum_i^{(1)}$	0.14242	3.74297	0.29849	4.04168
$\frac{4}{3}h\sum_i^{(1)}$	0.01898	0.49906	0.03980	0.53889
Y_{i-4}	3.14159	0.00000	0.14184	0.10981
$\bar{Y}_i$	3.16057	0.49906	3.18164	0.64870
Y'_{i-2}	0.03224	1.24773	0.08001	1.34734
$4Y'_{i-1}$	0.32004	5.38936	0.62804	5.7871
Y'_i	0.15698	1.44678	0.27079	1.54596
$\sum_i^{(2)}$	0.50926	8.08387	0.97884	8.68042
$\frac{h}{3}\sum_i^{(2)}$	0.01698	0.26945	0.03263	0.28935
Y_{i-1}	3.14364	0.22960	3.14903	0.35934
$\bar{\bar{Y}}_i$	3.16062	0.49905	3.18166	0.64809

and (10)]. Using then the data obtained and formula (3), we find $\bar{\bar{y}}_4$ and $\bar{z}_4$ (Table 9.18).

We transfer the values of $\bar{y}_4$ and $\bar{z}_4$ into the main table [columns (3) and (4)], determine $\bar{y}'_4$ and $\bar{z}'_4$ [columns (5) and (6) of Table 9.17] and then, using again the auxiliary table, we find $\bar{\bar{y}}_4$ and $\bar{\bar{z}}_4$. We transfer their values into columns (7) and (8) of Table 9.17. We find that

$$\varepsilon_{4,1} = \frac{1}{29}(\bar{\bar{y}}_4 - \bar{y}_4), \quad \varepsilon_{4,2} = \frac{1}{29}(\bar{\bar{z}}_4 - \bar{z}_4)$$

[columns (11) and (12) of Table 9.17]. For the preassigned permissible error $\varepsilon = 0.0001$ we see that we can omit the recalculation of the initial interval and take $\bar{\bar{y}}_4$ as y_4. To find $y\ (0.5)$ and $z\ (0.5)$, we repeat the whole process of calculation.

The final result is $y\ (0.4) = 3.16062$, $z\ (0.4) = 0.49905$, $y\ (0.5) = 3.18166$, $z\ (0.5) = 0.64869$. ▲

9.9. The Notion of the Boundary-Value Problem for Ordinary Differential Equations

We use the example of a second-order equation

$$F(x,\ y,\ y',\ y'') = 0 \tag{1}$$

to discuss the solution of the boundary-value problem for ordinary differential equations.

The simplest two-point boundary-value problem for equation (1) is posed as follows: we have to find the function $y = y\ (x)$ which satisfies equation (1) within the interval $[a,\ b]$ and the boundary conditions

$$\begin{aligned} \varphi_1\,[y\ (a),\ y'\ (a)] &= 0, \\ \varphi_2\,[y\ (b),\ y'\ (b)] &= 0 \end{aligned} \tag{2}$$

at its endpoints.

Let us consider some kinds of a two-point boundary-value problem for equation (1).

Assume, for instance, that we are given a second-order differential equation

$$y'' = f\ (x,\ y,\ y') \tag{3}$$

with boundary conditions $y\ (a) = A$, $y\ (b) = B$ $(a < b)$, i.e. the values of the required function $y = y\ (x)$ at the boundary points $x = a$ and $x = b$ are known. Then, in terms of geometry, the solution of equation (3) is an

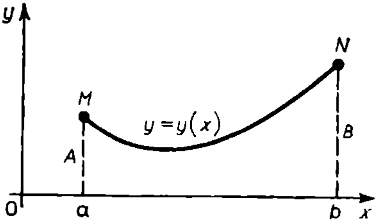


Fig. 9.4

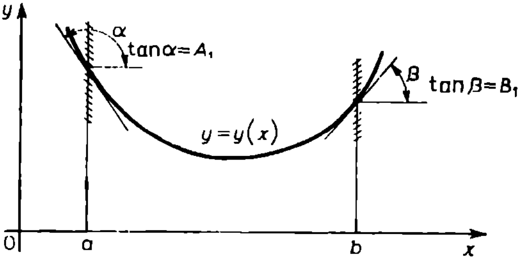


Fig. 9.5

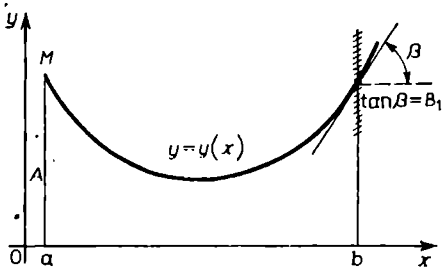


Fig. 9.6

integral curve $y = y(x)$ which passes through the given points $M(a, A)$ and $N(b, B)$ (Fig. 9.4).

Assume now that, for equation (3), we are given the values of the derivatives of the required function at the boundary points, i.e. $y'(a) = A_1$, $y'(b) = B_1$. Then, in terms of geometry, the solution of equation (3) means that we have to find an integral curve $y = y(x)$ of this equation which would cut the straight lines $x = a$ and $x = b$ at the angles $\alpha = \arctan A_1$ and $\beta = \arctan B_1$ respectively (Fig. 9.5).

Assume, finally, that, for equation (3), we know the value of the required function $y(a) = A$ at one boundary point and the value of the derivative of this function $y'(b) = B_1$ at the other point. A boundary-value problem of this kind is known as the *third* (*mixed*) boundary-value problem. In terms of geometry, the solution of equation (3) means that we have to find an integral curve $y = y(x)$ of this equation which would pass through the point $M(a, A)$ and cut the straight line $x = b$ at an angle $\beta = \arctan B_1$ (Fig. 9.6).

If the differential equation and the boundary conditions are linear, then the problem is *linear*. In that case the differential equation (1) and the boundary conditions (2) are written as

$$y'' + p(x)\,y' + q(x)\,y = f(x), \tag{4}$$

$$\begin{cases} \alpha_0 y(a) + \alpha_1 y'(a) = \gamma_1, \\ \beta_0 y(b) + \beta_1 y'(b) = \gamma_2, \end{cases} \tag{5}$$

where $p(x)$, $q(x)$, $f(x)$ are known functions continuous on the interval $[a, b]$, α_0, α_1, β_0, β_1, γ_1, γ_2 are given constants, and $|\alpha_0| + |\alpha_1| \neq 0$, $|\beta_0| + |\beta_1| \neq 0$.

If $f(x) = 0$ for $a \leqslant x \leqslant b$, then the equation is homogeneous otherwise it is *inhomogeneous*.

If $\gamma_1 = 0$ and $\gamma_2 = 0$, then the corresponding boundary condition is *homogeneous*. If both the differential equation and the boundary conditions are homogeneous, then the boundary-value problem is *homogeneous*.

9.10. The Method of Finite Differences for Second-Order Linear Differential Equations

Assume that we are given a second-order linear differential equation

$$y'' + p(x)\,y' + q(x)\,y = f(x) \tag{1}$$

with two-point linear boundary conditions

$$\begin{cases} \alpha_0 y(a) + \alpha_1 y'(a) = A, \\ \beta_0 y(b) + \beta_1 y'(b) = B \end{cases} \tag{2}$$

$$(|\alpha_0| + |\alpha_1| \neq 0, \quad |\beta_0| + |\beta_1| \neq 0)$$

and that $p(x)$, $q(x)$ and $f(x)$ are continuous on the interval $[a, b]$. Furthermore, assume that $x_0 = a$, $x_n = b$, $x_i = x_0 + ih$ $(i = 1, 2, \ldots, n-1)$ are systems of equispaced nodes with a step $h = (b-a)/n$ and $p_i = p(x_i)$, $q_i = q(x_i)$, $f_i = f(x_i)$.

We designate the approximate values of the function $y(x)$ and its derivatives $y'(x)$ and $y''(x)$ obtained as a result of calculations at the nodes x_i and y_i, y'_i and y''_i respectively. In each interior node we approximately replace the derivatives $y'(x_i)$ and $y''(x_i)$ by the finite-difference relations

$$y'_i = \frac{y_{i+1} - y_i}{h}, \quad y''_i = \frac{y_{i+2} - 2y_{i+1} + y_i}{h^2}, \tag{3}$$

and set

$$y'_0 = \frac{y_1 - y_0}{h}, \quad y_n = \frac{y_n - y_{n-1}}{h} \tag{4}$$

for the endpoints $x_0 = a$ and $x_n = b$.

Using formulas (3) and (4) we approximately replace equation (1) and the boundary conditions (2) by a system of equations

$$\begin{cases} \dfrac{y_{i+2} - 2y_{i+1} + y_i}{h^2} + p_i \dfrac{y_{i+1} - y_i}{h} + q_i y_i = f_i, \\ \alpha_0 y_0 + \alpha_1 \dfrac{y_1 - y_0}{h} = A, \quad \beta_0 y_n + \beta_1 \dfrac{y_n - y_{n-1}}{h} = B. \end{cases} \tag{5}$$

We thus arrive at an algebraic system of $n + 1$ equations in $n + 1$ unknowns. Solving such a system, we obtain a table of approximate values of the required function.

If we replace $y'(x_i)$ and $y''(x_i)$ by the central-difference relations

$$y'_i = \frac{y_{i+1} - y_{i-1}}{2h}, \quad y''_i = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2}, \tag{6}$$

we can get more accurate formulas. However, we can use formulas (5) for the derivatives at the endpoints. Then we get a system

$$\begin{cases} \dfrac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + p_i \dfrac{y_{i+1} - y_{i-1}}{2h} + q_i y_i = f_i, \\ \alpha_0 y_0 + \alpha_1 \dfrac{y_1 - y_0}{h} = A, \\ \beta_0 y_n + \beta_1 \dfrac{y_n - y_{n-1}}{h} = B. \end{cases} \tag{7}$$

Example. Using the finite-difference method, find the solution of the boundary-value problem

$$\begin{cases} y'' - xy' + 2y = x + 1, \\ y(0.9) - 0.5y'(0.9) = 2, \\ y(1.2) = 1 \end{cases}$$

with an accuracy of 0.001.

△ We divide the interval $[0.9, 1.2]$ into parts with a step $h = 0.1$. Then we get four nodal points with abscissas $x_0 = 0.9$, $x_1 = 1.0$, $x_2 = 1.1$, $x_3 = 1.2$. At the interior points $x_1 = 1.0$ and $x_2 = 1.1$ we replace this equation by the finite-difference equation

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} - x_i \frac{y_{i+1} - y_{i-1}}{2h} + 2y_i = x_{i+1} \; (i = 1, 2). \qquad (*)$$

Using the boundary conditions, we set up finite-diffference equations at the endpoints:

$$\begin{cases} y_0 + 0.5 \dfrac{y_1 - y_0}{h} = 2, \\ y_3 = 1. \end{cases} \qquad (**)$$

Collecting terms and taking into account that $n = 0.1$, we rewrite equations $(*)$ and $(**)$ in the forms

$$y_{i-1}(2 + 0.1x_i) - 4y_i(1 - 0.01) + y_{i+1}(2 - 0.1x_i) = 0.02(x_i + 1),$$
$$1.2y_0 - y_1 = 0.4, \; y_3 = 1$$

respectively.

This problem reduces to the solution of the system of equations

$$\begin{cases} 1.2y_0 - y_1 = 0.4, \\ 2.1y_0 - 3.96y_1 + 1.9y_2 = 0.04, \\ 1.11y_1 - 3.96y_2 + 1.89y_3 = 0.042. \end{cases}$$

Solving this system, we obtain $y_0 = 1.406$, $y_1 = 1.287$, $y_2 = 1.149$, $y_3 = 1.000$. ▲

Exercises

1. Using Picard's method, find three successive approximation of the solution of the following differential equations:

(a) $y' = 4y(1 + x)$; the initial condition is $y(0) = 1$,

(b) $y' = x - y$; the initial condition is $y(0) = 1$.

2. Find the first seven terms of the expansion in a power series of the solution $y = y(x)$ of the equation $y'' + 0.1(y')^2 + (1 + 0.1x)y = 0$ for the initial conditions $y(0) = 1$, $y'(0) = 2$.

3. Find a solution of the differential equation $y' = x^2 + y^2$ which would satisfy the initial condition $x_0 = 0$, $y(x_0) = 0$. Restrict the calculations to the terms of the expansion in a power series which involve x^7.

4. Setting $h = 0.1$, use Euler's method to solve the following differential equations for the given initial conditions on the indicated intervals:

(a) $y' = y + 3x$, $y(0) = -1$, $x \in [0, 0.5]$,
(b) $y' = x - 2y$, $y(0) = 0$, $x \in [0, 1]$.

5. Using improved Euler's method, find, on the interval $[0, 1]$, the table of solution of the differential equation $y' = y - \frac{2x}{y}$ for the initial condition $y(0) = 1$ assuming that $h = 0.2$.

6. Using the improved Euler-Cauchy method, solve the differential equation from Exercise 5.

7. Taking $h = 0.1$ find, by the Runge-Kutta method, the solutions of the following differential equations for the given initial conditions on the indicated intervals:
(a) $y' = x + y^2$, $y(1) = 0$, $x \in [1, 2]$,
(b) $y' = x^2 - y$, $y(0) = 2$, $x \in [0, 1]$.

8. Use Adams' extrapolation method to solve the differential equation $y' = 2x - y$ for the initial condition $y(0) = 1$ on the interval $[0, 1]$. The initial interval of the solution is given: $y_0 = 1$, $y_1 = 0.9145$, $y_2 = 0.8562$, $y_3 = 0.8225$ (assume that $h = 0.1$).

Chapter 10

Approximate Methods of Solution of Partial Differential Equations

10.1. Classification of the Second-Order Differential Equations

Many important practical problems of hydrodynamics, heat and mass transfer, heat conduction, diffusion, the theory of elasticity and other fields of knowledge are described by linear partial differential equations of the second order, among which, *equations in two independent variables** admit of the most simple and visual physical interpretation:

$$a_{11}u_{xx}+2a_{12}u_{xy}+a_{22}u_{yy}+b_1u_x+b_2u_y+cu=F, \tag{1}$$

where $u(x, y)$ is an unknown function which must be defined, a_{11}, a_{12}, a_{22}, b_1, b_2, c are specified functions of the independent variables x and y known as the *coefficients* of the equation, and F is the given function of x and y which is the *right-hand side* of the equation. If the coefficients of equation (1) are constant, then it is a *linear equation with constant coefficients*. Equation (1) is *homogeneous* if $F=0$.

A *solution* (*integral*) of equation (1) is any function which, being substituted for u in the equation, turns the equation into an identity.

The necessary conditions for the existence and uniqueness of equation (1) essentially depend on the coefficients, a_{11}, a_{12}, a_{22}. We shall assume that at least one of the coefficients is not identically zero (otherwise we would have a first-order equation). It turns out that the properties of the solution of equation (1) depend, to a large extent, on the value (the sign, to be more precise) of the *discriminant* $\Delta = a_{12}^2 - a_{11}a_{22}$.

---

* Here and henceforth we use the following designations for the derivatives: $u_x=\frac{\partial u}{\partial x}$, $u_{xx}=\frac{\partial^2 u}{\partial x^2}$, $u_{xy}=\frac{\partial^2 u}{\partial x\partial y}$ and so on.

In connection with the difference in the properties of solutions and, consequently, the methods of solution, the following classification of equations is accepted. Assume that the discriminant retains sign or is zero everywhere in a domain D.

Equation (1) is *elliptic* if $\Delta < 0$, *parabolic* if $\Delta = 0$ and *hyperbolic* if $\Delta > 0$.

If the discriminant changes sign when passing from one point of the domain D to another, then the equation is of a *mixed type*.

This classification is due to the fact that elliptic, parabolic and hyperbolic equations describe problems which are essentially different in their physical meaning and which deal with physical phenomena different in nature. Thus equations of heat conduction and diffusion (parabolic) express the laws of conservation of energy and matter. These equations are constructed on the basis of the laws of Fourier and Nernst which are similar from the point of view of mathematical formulation.

On the other hand, the equation of oscillation of a string (hyperbolic) is the law of conservation of momentum and is based on Newton's second law.

Finally, an elliptic equation specifies a function which is quite different in nature and which defines stationary processes which do not vary in time.

We know from the course in mathematical physics that under certain conditions imposed on the coefficients a_{11}, a_{12} and a_{22} (say, if they are twice continuously differentiable) there may occur a transformation of the variables

$$\xi = \varphi(x, y), \eta = \psi(x, y), \tag{2}$$

which turns equation (1) into one of the following canonical forms:

$$u_{\xi\xi} + u_{\eta\eta} = f \text{ (an elliptic equation)}, \tag{3}$$

$$u_{\xi\xi} = f \text{ (a parabolic equation)}, \tag{4}$$

$$u_{\xi\eta} = f \text{ or } u_{\xi\xi} - u_{\eta\eta} = f \text{ (a hyperbolic equation)}. \tag{5}$$

Here $f = f(\xi, \eta, u, u_\xi, u_\eta)$ is a function of independent variables, of an unknown function and its first derivatives. Note that a hyperbolic equation has two equivalent canonical forms.

For equation (1) with constant coefficients the transformation of the variables (2) is linear and simple in form. Consider these transformations for each type of the equation.

For elliptic equations

$$\xi = y - \frac{a_{12}}{a_{11}} x, \quad \eta = -\frac{\sqrt{a_{11}a_{22} - a_{12}^2}}{a_{11}} x. \tag{6}$$

For parabolic equations

$$\xi = y - \frac{a_{12}}{a_{11}} x, \quad \eta = x. \tag{7}$$

For hyperbolic equations

$$\xi = y - \frac{a_{12}}{a_{11}} x, \quad \eta = -\frac{\sqrt{a_{12}^2 - a_{11}a_{22}}}{a_{11}} x. \tag{8}$$

Note that for parabolic equations η may, in general, be an arbitrary function independent of ξ.

Example. Reduce the equation $u_{xx} + u_{xy} = F$ to canonical form.

△ Here $a_{11} = 1$, $a_{12} = 0.5$, $a_{22} = 0$ and, therefore, the discriminant $\Delta = 0.5^2 - 1 \cdot 0 = 0.25 > 0$. Consequently, this is a hyperbolic equation. Using the appropriate transformation of variables (8), we obtain

$$u_{xx} = \frac{1}{4}(u_{\xi\xi} + 2u_{\xi\eta} + u_{\eta\eta}), \quad u_{xy} = \frac{1}{2}(u_{\xi\xi} + u_{\eta\eta}).$$

Thus the canonical form of the original equation is

$$u_{\xi\xi} - u_{\eta\eta} = -4\overline{F}(\xi, \eta),$$

where

$$\overline{F}(\xi, \eta) = F(-2\eta, \xi - \eta).$$

We have obtained the second canonical form.

We can show that the transformation of the variables $\alpha = \xi + \eta$, $\beta = \xi - \eta$ leads to the first canonical form $u_{\alpha\beta} = -F(\beta - \alpha, \beta)$. ▲

In what follows, we consider the canonical forms (3)-(5) of equation (1).

10.2. Classification of Boundary-Value Problems

In this section we discuss the simplest canonical equations (3)-(5) from 10.1. We have mentioned (in Sec. 10.1) that different types of equations describe different physical processes. Thus, the equation*

$$u_{tt} - u_{xx} = f(x, t), \tag{1}$$

is the *equation of oscillation of a string* and describes the processes connected with the mechanical, electrical, acoustic and other kinds of oscillation.

The equation

$$u_t - u_{xx} = f(x, t), \tag{2}$$

known as the *equation of heat conduction*, describes heat flow, diffusion and other processes of transfer.

The equation

$$u_{xx} + u_{yy} = f(x, y), \tag{3}$$

known as *Poisson's equation* describes a stationary thermal field, a potential flow of fluid and other physical phenomena connected with transition to a steady state.

To describe a physical process completely (uniquely), we must indicate the equation of this process and, in addition, specify the initial conditions (the initial state of the process) and the conditions for variation of a certain function on the boundary of the domain in which the process takes place. In terms of mathematics, this is connected with the nonuniqueness of the solution of the differential equation. Therefore, to define the solution uniquely, we must define the equation itself and, in addition, to impose additional conditions which are classified as initial and boundary conditions. Three types of boundary-value problems are distinguished.

Initial conditions are conditions which define, at a certain moment, known as the initial moment, the value of the required solution (and sometimes its time derivatives too) for all points of the domains being considered.

Cauchy's problem is the first type of boundary-value problems (initial-value problem). This is the problem of solving equation (1) or (2) for which only initial conditions (the initial state of the process) are defined as additional conditions.

---

* Here and henceforth, for better physical visualness, the variables x and y correspond to the space coordinates and t to the time coordinate.

Cauchy's problem does not involve boundary conditions. This problem is formulated for hyperbolic and parabolic equations. The absence of boundary conditions is due to the fact that we consider either an unbounded domain or a small initial time interval when the effect of the boundary is negligibly small.

A *problem without initial conditions*, which involves only boundary conditions, is the second type of a boundary-value problem. In their turn, boundary conditions are usually divided into three kinds.

A ***boundary condition of the first kind*** is a condition under which the required function assumes specified values on the boundary of the domain being considered.

A ***boundary condition of the second kind*** is a condition under which the normal derivative of the required function must assume specified values on the boundary of the domain being considered.

A ***boundary condition of the third kind*** is a condition under which a linear combination of the required function and its normal derivative is specified on the boundary of the domain being considered.

From the point of view of mathematics, boundary conditions of the first and second kinds are special cases of conditions of the third kind. However, they have been separated not only for historical reasons but also due to their essentially different physical interpretation and a definite difference in the methods of solving the corresponding boundary-value problems.

The initial conditions in applied problems mav be absent when these problems involve moments of time sufficiently distant from the initial time moment, when the effect of the initial conditions is weak. Problems of this kind are often called *steady-state problems*. Problems of this type can be formulated for all kinds of equations (1)-(3).

The third kind of a boundary-value problem is a *mixed problem* in which both initial and boundary conditions are specified. This is a generalization, in a certain sense, of problems of the first two kinds. Cauchy's problem and a boundary-value problem without initial conditions are two extreme limiting cases of the mixed problem. The former is the limiting case for t sufficiently small time interval and the latter, for a sufficiently large time interval. The mixed problem is formulated for hyperbolic and parabolic equations.

10.3. Statement of the Simplest Boundary-Value Problems

In this section we consider various statements of boundary-value problems on the assumption that their solutions are sufficiently smooth. Here and in what follows we understand the sufficient smoothness of a function to be the continuity of the function and of the necessary number of its derivatives. In the statement of boundary-value problems, the sufficient smoothness usually means the continuity of all functions and derivatives appearing in the differential equation and the boundary conditions.

The *classical solution of a boundary-value problem* is any function which satisfies the differential equation at every interior point of the domain of definition of this equation and is continuous in the domain being considered, including the boundary.

The corresponding statement of a boundary-value problem is called *classical.* Thus a classical statement automatically imposes some restrictions on the input data of the boundary-value problem. Thus, for instance, it is required that the right-hand side of equations (1)-(3) be continuous and the boundary functions be sufficiently smooth. Note that in the most interesting applied problems the right-hand sides have essential peculiarities and therefore it is not sufficient to make a classical statement. In such cases we introduce the concept of a generalized solution, which we shall not discuss here but shall only consider the classical solution.

Let us pass now to the statement of problems.

Cauchy's problem for an unbounded domain. We shall formulate this problem for the equation of oscillation of a string and the equation of heat conduction.

We shall consider the process of oscillation of a thin infinite (very long) string under the action of a continuously distributed external force with density f. We assume that the force acts in one plane (Fig. 10.1), the plane of oscillation of the string (x, u) and the string is a flexible elastic thread. Let the value of the tension which appears in the string due to its bending obey Hooke's law and the oscillations themselves be sufficiently small. Then the value of the displacement $u(x, t)$ satisfies the equation of oscillation of the string

$$u_{tt} - u_{xx} = f(x, t) \quad (t > 0, \ -\infty < x < \infty). \qquad (1)$$

For the process to be unique, we must also specify the initial displacement and the initial velocity distribution. In terms of mathematics, this corresponds to the specification of the initial conditions:

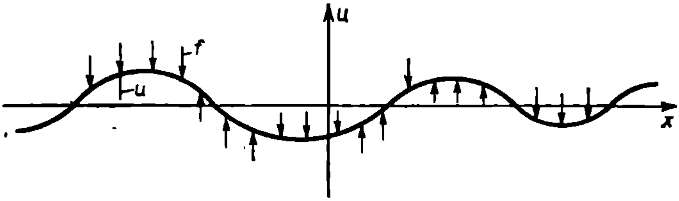


Fig. 10.1

$$u(x, 0) = u_0(x), \ u_t(x, 0) = u_1(x). \qquad (2)$$

We have to find the classical solution of equation (1) which satisfies the initial conditions (2).

Thus formulated, problem (1), (2) is known as *Cauchy's problem for a hyperbolic equation.*

Let us investigate now the process of temperature distribution in a thin infinite (very long) bar. We assume that the heat flow obeys Fourier's law and the variation of the body temperature is proportional to the quantity of heat imparted to the body. We assume that inside the bar the heat, characterized by the density of heat sources f can be liberated and absorbed. Then the distribution of temperature in the bar is described by the equation of heat conduction

$$u_t - u_{xx} = f(x, t) \quad (t > 0, -\infty < x < \infty). \qquad (3)$$

For a unique definition of the process, it is necessary to indicate the initial distribution of temperature. This corresponds to the specification of the initial condition

$$u(x, 0) = u_0(x). \qquad (4)$$

We have to find the classical solution of equation (3) which would satisfy the initial condition (4).

The problem (3), (4) we have formulated is known as *Cauchy's problem for parabolic equations.*

When posing boundary-value problems and especially when finding numerical solutions, it is necessary to answer the following three fundamental questions corresponding to the natural physical requirements:

(1) whether a solution of the problem exists and whether or not the boundary conditions redefine it,

(2) if a solution exists, then whether it is unique,

(3) whether the solution continuously depends on the initial data of the boundary-value problem (f, u_0, u_1, etc.), i.e. whether it varies continuously with a continuous variation of the right-hand side of the equation and the boundary conditions. This property is called the *stability of the solution* relative to the input data.

A boundary-value problem is *correct* if its solution exists, is unique and stable.

The classical Cauchy problem for the equation of oscillation of a string is correct if the functions f, u_0 and u_1 are sufficiently smooth.

For Cauchy's problem for the equation of heat conduction to be correct, in addition to the smoothness of f and u_0, it is also necessary that the solution be limited.

A stationary problem (a problem without initial data). Let us consider a steady-state condition of temperature distribution in a bounded thin plate of an arbitrary shape with a smooth boundary. Let the function $u(x, y)$ express the temperature at every point of the plate. Under the ordinary laws of heat distribution described above when Cauchy's problem for the equation of heat conduction was formulated, the function $u(x, y)$ satisfies Poisson's equation

$$u_{xx} + u_{yy} = f(x, y), \quad (x, y) \in D, \tag{5}$$

where the function f defines the density of heat sources of the plate. When there are no sources ($f = 0$), equation (5) becomes *Laplace's equation*:

$$u_{xx} + u_{yy} = 0. \tag{6}$$

For the description of the process to be unique, we must define the heat conditions on the boundary of the plate. We can do this by specifying or distributing the temperature on the boundary, or, else, by distributing the heat flow.

The conditions of heat balance of the radiating body and the medium are also possible. Depending on heat conditions on the boundary, we distinguish three boundary conditions for the function $u(x, y)$. Let Γ be the boundary of the domain D of definition of equation (6). In terms of mathematics, the boundary conditions can be formulated as follows:

the boundary condition of the first kind:

$$u|_\Gamma = \varphi_0(x, y), \quad (x, y) \in \Gamma, \tag{7}$$

the boundary condition of the second kind:

$$\left.\frac{\partial u}{\partial n}\right|_\Gamma = \varphi_1(x, y), \ (x, y) \in \Gamma, \tag{8}$$

the boundary condition of the third kind:

$$\left.\frac{\partial u}{\partial n}\right|_\Gamma + \lambda u\Big|_\Gamma = \varphi_2(x, y) \quad (x, y) \in \Gamma. \tag{9}$$

The derivative is taken with respect to the outer normal to the curve Γ, $\lambda > 0$ is the thermal conductivity, φ_0, φ_1 and φ_2 are functions defined on Γ, φ_2 being the product of the thermal conductivity by the temperature of the medium which is in contact with the body.

Thus the boundary-value problem consists in finding the classical solution of equation (5) or (6) satisfying one of the boundary conditions (7)-(9).

Depending on the kind of the boundary conditions, three boundary-value problems are distinguished: the first boundary-value problem (5), (7), which is *Dirichlet's problem*, the second boundary-value problem (5), (8) which is *Neumann's problem* and the third boundary-value problem (5), (9).

For sufficiently smooth input data, the first and the third problem for the equations of Poisson and Laplace are correct.

For Neumann's problem, the uniqueness theorem consists in the fact that under the same boundary conditions two of its solutions may vary by a constant quantity.

In problems of mathematical physics, of considerable importance are harmonic functions.

A function $u(x, y)$ is *harmonic* in a domain D if it is continuous together with its second-order derivatives and satisfies equation (6) in this domain.

Here are some properties of harmonic functions.

1° **(principle of the maximum).** *If the function* $u(x, y)$ *is defined and continuous in* $\bar{D} = D \cup \Gamma$ *and satisfies equation* (6) *in* D, *then* u *attains its maximum and minimum values on the boundary* Γ, i.e.

$$\max_{(x,\, y) \in D} u(x, y) \leqslant \max_{(x,\, y) \in \Gamma} u(x, y),$$

$$\min_{(x,\, y) \in D} u(x, y) \geqslant \min_{(x,\, y) \in \Gamma} u(x, y).$$

2° **(a consequence of property 1°).** *If the functions* $u(x, y)$ *and* $v(x, y)$ *are continuous in* $\bar{D}$, *harmonic in* D *and* $u \leqslant v$ *for* $(x, y) \in \Gamma$, *then* $u \leqslant v$ *for* $(x, y) \in D$.

3° **(a consequence of property 1°).** *If the functions* u *and* v *are continuous in* $\bar{D}$, *harmonic in* D *and* $|u| \leqslant v$ *for* $(x, y) \in \Gamma$ *then* $|u| \leqslant v$ *for* $(x, y) \in D$.

The mixed boundary-value problem. Let us consider a problem of heat propagation in a thin bar of unit length. We place one of its ends at the point $x = 0$ and the other end at the point $x = 1$. In the time interval $0 < t < T$ the temperature distribution in such a bar is described by the equation

$$u_t - u_{xx} = f \quad (0 < x < 1,\ 0 < t \leqslant T) \tag{10}$$

with the initial condition

$$u(x, 0) = u_0(x) \quad (0 \leqslant x \leqslant 1), \tag{11}$$

and for the solution to be unique in this case, it is also necessary to specify the temperature conditions at the ends of the bar. We can do this with the aid of boundary conditions similar to those formulated for the equations of Poisson and Laplace.

The boundary condition of the first kind (the temperature is specified at the end of the bar $x = 0$) is

$$u(0, t) = \varphi_0(t) \quad (0 < t \leqslant T). \tag{12}$$

The boundary condition of the second kind (the heat flow is specified at the end of the bar $x = 0$) is

$$u_x(0, t) = \varphi_1(t) \quad (0 < t \leqslant T). \tag{13}$$

The boundary condition of the third kind is

$$-u_x(0, t) + \lambda u(0, t) = \varphi_2(t) \quad (0 < t \leqslant T). \tag{14}$$

For the other end of the bar $x = 1$, the right-hand sides of the boundary conditions (12)-(14) are replaced by $\psi_0(t)$, $\psi_1(t)$ and $\psi_2(t)$ respectively. Note that the initial and the boundary condition must satisfy the so-called *matching conditions*, i.e. $u_0(0) = \varphi_0(0)$ under condition (12), $u_{0x}(0) = \varphi_1(0)$ under condition (13) and $-u_{0x}(0) + \lambda u_0(0) = \varphi_2(0)$ under condition (14). Similar matching conditions must be fulfilled at the other end of the bar $x = 1$.

Thus, for the first boundary-value problem the matching conditions mean that

$$u_0(0) = \varphi_0(0), \; u_0(1) = \psi_0(0). \tag{15}$$

In the general case, different conditions may be at different ends of the bar so that the total number of all possible combinations of boundary conditions is 6.

Here is one of the possible boundary-value problems. We have to find a solution of equation (10) which would satisfy the initial condition (11) and the following boundary conditions:

$$u(0, t) = \varphi_0(t), \; u(1, t) = \psi_0(t) \; (0 < t \leqslant T). \tag{16}$$

Problem (10), (11), (16) is known as the first boundary-value problem for the equation of heat conduction. Correspondingly, the boundary-value problem (10), (11) with the boundary conditions (13) or (14) at both ends of the bar is the second or the third problem.

Other boundary-value problems with various combinations of boundary conditions (12)-(14) at both ends of the bar are posed by analogy.

Boundary-value problems of the first, the second and the third kind are correct if the corresponding conditions of smoothness and matching are fulfilled for the input data.

The solutions of the equation of heat conduction possess the following significant property, which is similar to the property presented for the solution of Laplace's equation.

The principle of the maximum. *If the function* $u(x, t)$ *is continuous in the domain* $\bar{D}_T$ $\{0 \leqslant t \leqslant T, 0 \leqslant x \leqslant 1\}$ *and satisfies equation* (10), *then for* $f = 0$, *the maximum and minimum values of the function* $u(x, t)$ *are attained either at the initial moment or at the point* $x = 0$ *or* $x = 1$ *of the boundary.*

Let us consider now the oscillation of a thin string of unit length. The value of the displacement $u(x, t)$ is described by the hyperbolic equation

$$u_{tt} - u_{xx} = f \quad (0 < x < 1, t > 0). \qquad (17)$$

In this case the initial conditions have the form

$$u(x, 0) = u_0(x), u_t(x, 0) = u_1(x) \quad (0 \leqslant x \leqslant 1). \quad (18)$$

We shall consider the boundary conditions in the same form as for the equation of heat conduction, i.e. (12)-(14). Problem (17)-(18) with the same boundary conditions at both ends of form (12)-(14) is the first, the second and the third boundary-value problem, respectively, for a hyperbolic equation. All these problems are correct if the corresponding conditions of smoothness and matching are fulfilled for the input data.

As an example we shall formulate, the first boundary-value problem for the equation of oscillation of a string.

We have to find a function $u(x, t)$ which would satisfy equation (17), the initial conditions (18) and the following boundary conditions:

$$u(0, t) = \varphi_0(t), u(1, t) = \psi_0(t) \quad (t > 0). \qquad (19)$$

The physical meaning of the boundary conditions of the first kind is that both ends of the string oscillate in present modes according to the laws of the given functions φ_0 and ψ_0. The boundary condition of the second kind corresponds to the fact that a law of the action of a force is defined at the endpoint. The boundary condition of the third kind corresponds to the elastic fastening of the end of the string.

10.4. The Method of Finite Differences. The Principal Concepts

For the simplest boundary-value problems formulated in 10.3, certain exact solutions are given in various courses in mathematical physics. But the majority of even linear equations which describe practical processes are such that they do not admit of constructing an exact solution by means of elementary functions. In such cases we resort to approximate methods. We usually consider two kinds of approximate solutions, analytical and numerical. We shall consider numerical methods based on the difference approximation of the derivatives. This approach is known as the **difference method**, or the **method of finite differences.**

To cut down the computations, we shall illustrate this method using the simplest equations (for which, maybe, an exact solution has been obtained), bearing in mind that the main principles of constructing difference schemes can be extended to more general equations.

Consider a linear differential equation written in the following symbolic form:

$$Lu(x, y) = f(x, y), \quad (x, y) \in D. \tag{1}$$

Here u is the required solution of the equation, L is a differential operator symbolizing the corresponding operation of differentiation, and f is the right-hand side of the equation (the given function).

As is known, for the solution of equation (1) to be unique, it must be supplemented with the boundary and initial conditions. We write these conditions in the form of a symbolic relation

$$lu(x, y) = \varphi(x, y), \quad (x, y) \in \Gamma, \tag{2}$$

where l is an operator symbolizing the left-hand side of the boundary condition, φ is the right-hand side of the boundary condition (the given function) and Γ is the boundary of the domain D.

We shall illustrate the principal concepts of the method of finite differences by the solution of the Dirichlet problem for the Laplace equation in the square D^0 $\{0 < x < 1, 0 < y < 1\}$ with the boundary Γ^0 $\{x =$

$0, x = 1, 0 \leqslant y \leqslant 1;\ y = 0, y = 1,\ 0 \leqslant x \leqslant 1\}$:

$$Lu \equiv u_{xx} + u_{yy} = 0,\ (x, y) \in D^0, \tag{3}$$

$$u\ (x, y) = \frac{1}{4} xy\ (x + 1)\ (y + 1),\ (x, y) \in \Gamma^0. \tag{4}$$

Thus, for problem (3), (4), the operator L transforms the function u into a differential expression $u_{xx} + u_{yy}$. In such cases we write that

$$L \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2},$$

the right-hand side of the equation $f = 0$, the operator of the boundary conditions is an identity operator, i.e. it transforms the function u into u: $lu \equiv u$, the right-hand side of the boundary condition has the form

$$\varphi(x, y) = \begin{cases} 0, & x = 0,\ 0 \leqslant y \leqslant 1, \\ 0, & 0 \leqslant x \leqslant 1,\ y = 0, \\ \frac{1}{2} y\,(y+1), & x = 1,\ 0 \leqslant y \leqslant 1, \\ \frac{1}{2} x\,(x+1), & 0 \leqslant x \leqslant 1,\ y = 1. \end{cases}$$

The difference method of solving problem (1), (2) can be presented as two stages:

(1) constructing the difference scheme which approximates the given continuous problem,

(2) solving the difference problem and estimating the error of the solution.

We shall dwell on these problems in more detail.

When constructing the difference scheme, the first step is to replace the domain $\overline{D}$ of the continuous variation of the arguments by the domain of their discrete variation, i.e. by the *net domain* (or simply *net*) $\overline{\omega}_h$, i.e. by the set of points (x_n, y_m) called the *nodal points*, or *nodes*, of the net. For the square $\overline{D}^0$ the net domain can be constructed as follows. We draw straight lines

$$x_n = nh,\ y_m = mh\ (h = 1/N,\ n, m = 0, 1, \ldots, N). \tag{5}$$

The set of intersection points (x_n, y_m) of these straight lines constitutes the net domain $\bar{\omega}_h$ and the points themselves form the *nodes* of the net. Every function $v(x, y)$, defined on the net $\bar{\omega}_h$, is a ***net function*** and is often designated as v_h.

The second step in constructing the difference scheme is the approximation (approximate replacement) of the differential expression Lu by a difference expression and the function f of the continuous argument by a net function, i.e in constructing a difference analogue for equation (1). The same refers to the boundary conditions (2).

Such an approximation leads to a system of algebraic equations with respect to the values of the net function v_h. We can write this system of equations as

$$L_h v_h = f_h, \tag{6}$$

$$l_h v_h = \varphi_h, \tag{7}$$

where L_h and φ_h are difference operators* which approximate L and l respectively, v_h is the required net function which approximates the solution u, f_h and φ_h are the given net functions which approximate f and φ respectively.

The collection of the difference equations (6), (7) which approximate the original problem (1), (2) is known as the *difference scheme*. Note that, in general, the original problem can be approximated by different difference schemes and one and the same difference scheme can approximate different continuous problems.

By way of example, we construct a difference scheme for problem (3), (4).

For the solution $u(x, y)$ we construct a net function u_h defined as follows: $u_h(x_n, y_m) = u(x_n, y_m)$. In what follows, in order to simplify the notation (where no ambiguity arises), we shall omit the index h in the net functions and especially in their values. Thus $u_{nm} = u_h(x_n, y_m)$. Using this notation, we approximate every derivative from equation (3) by a difference relation:

---

* Note that we understand the concept of the operator here as the abbreviation of the symbolic notation for differential and difference expressions,

$$u_{xx}(x_n, y_m) \cong \frac{1}{h^2}(u_{n-1,m} - 2u_{nm} + u_{n+1,m}),$$
$$u_{yy}(x_n, y_m) \cong \frac{1}{h^2}(u_{n,m-1} - 2u_{nm} + u_{n,m+1}). \tag{8}$$

Then we can approximate the differential equation (3) by the difference equations

$$\frac{1}{h^2}(v_{n-1,m} + v_{n,m-1} + v_{n+1,m} + v_{n,m+1} - 4v_{nm}) = 0 \tag{9}$$
$$(n,\ m = 1,\ 2,\ \ldots,\ N-1).$$

Hence we have

$$v_{nm} = \frac{1}{4}(v_{n-1,m} + v_{n+1,m} + v_{n,m-1} + v_{n,m+1}) \tag{10}$$
$$(n,\ m = 1,\ 2,\ \ldots,\ N-1).$$

The boundary condition (4) can be approximated as follows:

$$v_{0m} = 0,\ v_{n0} = 0,$$
$$v_{Nm} = \frac{1}{2}\,\frac{m(m+N)}{N^2},\quad v_{nN} = \frac{1}{2}\,\frac{n(n+N)}{N^2}$$
$$(n,\ m = 0,\ 1,\ \ldots,\ N). \tag{11}$$

The system of equations (10), (11) is usually solved by the method of simple iteration or by Seidel's method.

Comparing relations (6), (7) and (9), (11), it is easy to understand the meaning of the difference operators and net functions of system (6), (7) and their relations with the corresponding operators and functions of problem (1), (2).

The difference expression $L_h v_h$ is a linear combination of the values of the net function at some nodes. In particular, the difference expression (9) contains five nodes (a five-node scheme) known as a "cross" scheme (Fig. 10.2).

Thus for $(N-1)^2$ unknown values v_{nm} $(n, m = 1, 2, \ldots, N-1)$ of the net function v_h we get a system of $(N-1)^2$ equations (9) or (10) in which the quantities v_{0m}, v_{n0}, v_{Nm}, v_{nN} are defined by the boundary conditions (11). We can regard relations (9), (11) as a consistent system of $(N+1)^2-4$ equations in $(N+1)^2-4$ unknowns v_{nm} $(n, m=0, 1, \ldots, N)$, except for $v_{00}, v_{0N}, v_{NN}, v_{N0}$.

Intuition prompts us that the more accurate the approximation of type (8), the closer the function v_h to u_h. Therefore, at this stage we introduce a strict concept of approximation.

The solution v_h of problem (6), (7) is a net function and, consequently, it depends on the parameter h, i.e. on the

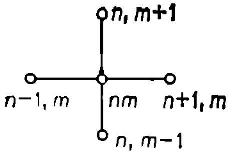


Fig. 10.2

step of the net. A natural question arises concerning the possibility, in principle, to approximate the solution $u(x, y)$ of problem (1), (2) by the solution v_h in a finite number of actions by choosing an appropriate stepsize h.

We shall compare two net functions v_h and $u_h = u(x, y)$, $(x, y) \in \bar{\omega}_h$. To find how close the two net functions are, we define the concept of the *norm* on the set of net functions as follows:

$$\|v_h\| = \max_{(x, y) \in \bar{\omega}_h} |v(x, y)|. \tag{12}$$

In definition (12) we take the maximum over the domain of definition of the function which is under the sign of the norm.

Let us consider the error of the difference scheme (6), (7): $z_h = v_h - u_h$. Substituting $v_h = z_h + u_h$ into (6), (7), we get, for z_h, a problem similar to that for v_h:

$$L_h z_h = f_h - L_h u_h, \tag{13}$$

$$l_h z_h = \varphi_h - l_h u_h. \tag{14}$$

The right-hand sides of equations (13), (14) constitute the *error of approximation of equation* (1) by the difference equation (6) and the *error of approximation of the boundary condition* (2) by the difference condition (7) on the solution of the original problem (1), (2).

We say that the difference scheme (6), (7) *approximates problem* (1), (2) *with the order* $k > 0$ *relative to* h on the solution $u(x, y)$ if

$$\|f_h - L_h u_h\| \leqslant c_1 h^k, \quad \|\varphi_h - l_h u_h\| \leqslant c_2 h^k, \tag{15}$$

where c_1 and c_2 are constants independent of h.

We shall determine the order of approximation for the scheme (9), (11), or, which is the same, (10), (11). Since the boundary conditions are exactly defined on the net, the left-hand side of the second inequality (15) is zero and the order of approximation is defined by the first inequality (15).

The right-hand side of equation (3) is zero and therefore we only estimate the norm of $L_h u_h$.

We use Taylor's formula:

$$u_{n\pm1,\,m} = u_{nm} \pm h\left(\frac{\partial u}{\partial x}\right)_{nm} + \frac{h^2}{2}\left(\frac{\partial^2 u}{\partial x^2}\right)_{nm} \pm \frac{h^3}{6}\left(\frac{\partial^3 u}{\partial x^3}\right)_{nm} + \frac{h^4}{24}\left(\frac{\partial^4 u}{\partial x^4}\right)^{\pm}_{xy},$$

$$\left(\frac{\partial^4 u}{\partial x^4}\right)^{\pm}_{xy} = \frac{\partial^4}{\partial x^4} u(x_n \pm \theta_1 h,\ y_m),\quad 0 < \theta_1 < 1,$$

$$u_{n,m\pm1} = u_{nm} \pm h\left(\frac{\partial u}{\partial y}\right)_{nm} + \frac{h^2}{2}\left(\frac{\partial^2 u}{\partial y^2}\right)_{nm} \pm \frac{h^3}{6}\left(\frac{\partial^3 u}{\partial y^3}\right)_{nm} + \frac{h^4}{24}\left(\frac{\partial^4 u}{\partial y^4}\right)^{\pm}_{xy},$$

$$\left(\frac{\partial^4 u}{\partial y^4}\right)^{\pm}_{xy} = \frac{\partial^4}{\partial y^4} u(x_n,\ y_m \pm \theta_2 h),\quad 0 < \theta_2 < 1.$$

Summing up these four relations, we obtain

$$u_{n+1,m} + u_{n-1,m} + u_{n,m+1} + u_{n,m-1} = 4u_{nm} + h^2\left[(u_{xx})_{nm} + (u_{yy})_{nm}\right] + \frac{h^4}{24}\left[\left(\frac{\partial^4 u}{\partial x^4}\right)^{+}_{xy} + \left(\frac{\partial^4 u}{\partial x^4}\right)^{-}_{xy} + \left(\frac{\partial^4 u}{\partial y^4}\right)^{+}_{xy} + \left(\frac{\partial^4 u}{\partial y^4}\right)^{-}_{xy}\right].$$

Using the difference relations (9), we find that

$$\|L_h u_h\| \leqslant \frac{h^2}{6} M_4, \tag{16}$$

here $M_4 = \max\limits_{(x,\, y)\in\overline{D}} \left(\left| \frac{\partial^4 u}{\partial x^4} \right|, \left| \frac{\partial^4 u}{\partial y^4} \right| \right)$. We assume here the existence of the corresponding derivatives. Thus scheme (9), (11) has the second order of approximation.

This completes the discussion of the first stage, i.e. the construction of a difference scheme, and now we pass to the second stage, the solution of the difference problem and evaluation of the error.

The most significant problems that arise at this stage are those of the solvability of the difference problem, the uniqueness of its solution and the continuous dependence of the solution on the input data. The *input data* are the right-hand sides of the difference equations and the boundary and initial conditions of the difference problem. We can formulate the concept of the correctness of a difference scheme in the same way as we pose the question concerning the correctness of problems of mathematical physics. Let v_h be a solution and f_h and φ_h the input data of a difference problem (6), (7). The solution and the input data evidently depend on h.

We say that a *difference problem* (*scheme*) *is correct* if the following conditions are fulfilled for all $h < h_0$ ($h_0 > 0$):

(1°) a solution of the difference problem exists and is unique,

(2°) the solution of the difference problem continuously depends on the input data.

We can write condition 2° for the difference problem (6), (7) as follows:

$$\| \overline{v}_h - v_h \| \leqslant M_1 \| \overline{f}_h - f_h \| + M_2 \| \overline{\varphi}_h - \varphi_h \|, \quad (17)$$

where the symbols without a bar correspond to one problem and those with a bar to the other.

This condition is known as the *stability of a difference problem* (*scheme*) *with respect to the input data*, or simply the *stability*.

It is proved in the theory of difference schemes that the "cross" scheme, constructed for Dirichlet's problem, is correct for Laplace's equation (and, in general, for Poisson's equation).

Having formulated the concepts of approximation and stability for difference schemes, we arrive at the most significant problem, that of the convergence of the solution of the difference problem (6), (7) to the solution of the continuous problem (1), (2).

We say that the difference scheme (6), (7) *converges at the rate of* $s > 0$ *relative to* h if the condition

$$\| v_h - u_h \| < ch^s$$

is satisfied, where c is a constant independent of h.

There is a close relationship between the concepts of approximation, correctness and stability which is defined by the following theorem.

Theorem. *Let the difference problem* (6), (7) *approximate problem* (1), (2) *on the solution* $u(x, y)$ *with the order of* $k > 0$ *relative to* h *and be correct. Then this scheme converges with the order equal to the order of approximation* k, *i.e. the estimate*

$$\| v_h - u_h \| \leqslant ch^k \tag{18}$$

is satisfied.

□ By the definition of approximation we have

$$\|f_h - L_h u_h\| \leqslant c_1 h^k, \quad \|\varphi_h - l_h u_h\| \leqslant c_2 h^k.$$

Using relations (13) and (14), we obtain

$$\| L_h z_h \| \leqslant c_1 h^k, \quad \| l_h z_h \| \leqslant c_2 h^k.$$

Then, by virtue of the assumption of the stability for the difference scheme [relation (17)], we have

$$\| z_h \| < M_1 \| L_h z_h \| + M_2 \| l_h z_h \| ,$$

whence, using the estimate just obtained, we find that

$$\| z_h \| = \| v_h - u_h \| \leqslant M_1 c_1 h^k + M_2 c_2 h^k = ch^k. \quad \blacksquare$$

Example 1. Find the solution of problems (3), (4):

$$u_{xx} + u_{yy} = 0, \quad (x, y) \in D^0,$$

$$u(x, y) = \frac{1}{4} xy (x+1)(y+1), \quad (x, y) \in \Gamma^0.$$

Here D^0 is a square $\{0 < x < 1, 0 < y < 1\}$ with the boundary Γ^0 $\{x = 0, x = 1, 0 \leqslant y \leqslant 1, y = 0, y = 1, 0 \leqslant x \leqslant 1\}$.

△ The system of the finite-difference equations has been written for this problem in the general case and has the form

$$v_{nm}=\frac{1}{2}(v_{n-1,\,m}+v_{n+1,\,m}+v_{n,\,m-1}+v_{n,\,m+1})$$

$$(n;\ m=1,\ 2,\ \ldots,\ N-1),$$

$$(v_{0m}=v_{n0}=0;$$

$$v_{Nm}=\frac{1}{2}\cdot\frac{m(m+N)}{N^2},\quad v_{nN}=\frac{1}{2}\cdot\frac{n(n+N)}{N^2}$$

$$(n;\ m=0,\ 1,\ \ldots,\ N).$$

Assuming $h = 1/3$ ($N = 3$) as the stepsize of the net, we construct a table of boundary conditions and unknown values:

Table 10.1

	0.222	0.556	
0	v_{12}	v_{22}	0.556
0	v_{11}	v_{21}	0.222
	0	0	

The initial system of equations for the unknown values assumes the form

$$v_{11}=\frac{1}{4}(0+v_{21}+0+v_{12})$$

$$v_{12}=\frac{1}{4}(0+v_{22}+v_{11}+0.222),$$

$$v_{21}=\frac{1}{4}(v_{11}+0.222+0+v_{22}),$$

$$v_{22}=\frac{1}{4}(v_{12}+0.556+v_{21}+0.556).$$

Let us now use the method of simple iteration to solve this system. To do this, we have to get the initial values of the unknowns. We shall obtain them by means of a linear interpolation using the boundary conditions, first for the rows and then for the columns.

We shall carry out the linear interpolation for the rows using the formula

$$\overline{v}_{nm}=v_{0m}+(v_{Nm}-v_{0m})\frac{m}{N}.$$

This yields (n, m = 1, 2)

$\overline{v}_{12}$ = 0.185, $\overline{v}_{22}$ = 0.371, $\overline{v}_{11}$ = 0.074, $\overline{v}_{21}$ = 0.148, We shall carry out the linear interpolation for the columns using the formula

$$\bar{\bar{v}}_{nm}=v_{n0}+(v_{nN}-v_{n0})\frac{n}{N}.$$

This yields (n, m, = 1, 2)

$$\bar{v}_{12}=0.148,\quad \bar{v}_{11}=0.074,\quad \bar{v}_{22}=0.371,\quad \bar{v}_{21}=0.185.$$

We take the half-sum of the values obtained

$$\overset{0}{v}_{nm}=\frac{1}{2}(\bar{v}_{nm}+\bar{\bar{v}}_{nm}),$$

i.e. $v_{12}^0=v_{21}^0=0.166$, $v_{22}^0=0.371$, $v_{11}^0=0.074$, as the initial values.

We can now perform the iteration process:

$$v_{11}^{k+1}=\frac{1}{2}v_{12}^k,\quad v_{12}^{k+1}=\frac{1}{4}(v_{11}^k+v_{22}^k+0.222),$$

$$v_{22}^{k+1}=\frac{1}{2}(v_{12}^k+0.556).$$

We have used the symmetry of the initial data ($v_{12}^0=v_{21}^0$) and of the system of equations. We shall continue with the solution of this system until two successive iterations coincide with an accuracy of 0.001. The results of calculations are in Table 10.2.

Table 10.2

	Number of iteration		
	0	1	2
v_{11} $v_{12}=v_{21}$ v_{22}	0.074 0.166 0.371	0.083 0.167 0.361	0.084 0.166 0.362

Two iterations proved to be sufficient to get a solution. This speaks of the simplicity of the difference problem due to the large stepsize of the net. ▲

Example 2. Find a solution of Laplace's equation (3) in a unit square under the following boundary conditions:

$$u(x,y)=\begin{cases}0, & 0\leqslant x\leqslant 1,\ y=0,\\ \frac{8}{3}y(64y^2-60y+29), & x=0,\ 0\leqslant y\leqslant 1,\\ \frac{8}{3}(1-x)(64x^2-68x+33), & 0\leqslant x\leqslant 1,\ y=1,\\ 0 & x=1,\ 0\leqslant y\leqslant 1,\end{cases}$$

△ We set $h = 0.25$ and construct system (10) taking the boundary conditions into account:

$$v_{11}^{k+1} = \frac{1}{4}(12 + v_{21}^k + 0 + v_{12}^k),$$

$$v_{21}^{k+1} = \frac{1}{4}\left(v_{11}^k + v_{31}^k + 0 + v_{22}^k\right),$$

$$v_{31}^{k+1} = \frac{1}{4}\left(v_{21}^k + 0 + 0 + v_{32}^k\right) = \frac{1}{2} v_{21}^k,$$

$$v_{12}^{k+1} = \frac{1}{4}\left(20 + v_{22}^k + v_{11}^k + v_{13}^k\right),$$

$$v_{22}^{k+1} = \frac{1}{4}\left(v_{12}^k + v_{32}^k + v_{23}^k + v_{21}^k\right) = \frac{1}{2}\left(v_{12}^k + v_{21}^k\right),$$

$$v_{13}^{k+1} = \frac{1}{4}\left(40 + v_{23}^k + v_{12}^k + 40\right) = 20 + \frac{1}{2} v_{12}^k,$$

The boundary conditions and the unknown values are given in Table 10.3.

Table 10.3

	40	20	12	
40	v_{13}	v_{23}	v_{33}	0
20	v_{12}	v_{22}	v_{32}	0
12	v_{11}	v_{21}	v_{31}	0
	0	0	0	

We have used the symmetry property, $v_{nm} = v_{N-m, N-n}$, to construct this system.

We shall calculate the initial approximation with the aid of the linear interpolation using the boundary conditions at the interior nodes. Using the formula

$$v_{n1}^0 = 12\left(1 - \frac{n}{4}\right)$$

to calculate v_{n1}^0, we get $v_{11}^0 = 9$, $v_{21}^0 = 6$, $v_{31}^0 = 3$. By virtue of symmetry, we set $v_{32}^0 = v_{21}^0 = 6$, $v_{33}^0 = v_{11}^0 = 9$.

Using the formula

$$v_{n2}^0 = 20\left(1 - \frac{7}{30} n\right)$$

to calculate v_{12}^0 and v_{22}^0, we get $v_{12}^0 = 15.33$, $v_{22}^0 = 10.66$. By virtue of symetry we set $v_{23}^0 = v_{12}^0 = 15.33$.

The last value, v_{13}^0, we get from the formula

$$v_{13}^0 = +40 - \frac{40-15.33}{2} \cdot 1 = 27.67.$$

We shall use two methods to solve this system: the method of simple iteration (Table 10.4) and Seidel's method (Table 10.5). We shall continue the calculations until two successive solutions for each variable coincide with an accuracy of 0.1. The method of simple iteration required four iterations and Seidel's method required three.

The final solutions are given in Tables 10.4 and 10.5. ▲

Table 10.4

	40	20	12	
40	28.5	17.0	8.6	0
20	17.0	11.3	5.6	0
12	8.6	5.6	2.8	0
	0	0	0	

Table 10.5

	40	20	12	
40	28.6	17.0	8.6	0
20	17.0	11.4	5.7	0
12	8.6	5.7	2.8	0
	0	0	0	

10.5. Difference Schemes for Solving the Equation of Heat Conduction

Let us consider the first boundary-value problem for the equation of heat conduction in a rectangle $\bar{D}$ $\{0 \leqslant x \leqslant 1,\ 0 \leqslant t \leqslant T\}$. We have to find a solution of the problem

$$Lu \equiv u_t - u_{xx} = f \quad (0 < x < 1,\ 0 < t \leqslant T), \qquad (1)$$

$$u(x,\ 0) \equiv u_0(x) \qquad (0 \leqslant x \leqslant 1), \qquad (2)$$

$$u(0,\ t) = \varphi_0(t),\ u(1,\ t) = \psi_0(t)\ (0 \leqslant t \leqslant T) \qquad (3)$$

continuous in $\bar{D}$.

As we did in 10.4 for Poisson's equation, we use the difference method to construct the solution of problem (1)-(3).

In the domain $\bar{D}$, we introduce a uniform rectangular net $\bar{\omega}_{h\tau}\ \{x_n,\ t_k\}$ with the stepsize $h = 1/N$ with respect to the coordinate x and with the stepsize $\tau = T/M$ with respect to the coordinate t:

$$x_n = nh\ (n = 0,\ 1,\ \ldots,\ N),\ t_k = k\tau\ (k = 0,\ 1,\ \ldots\ M).$$

We approximate the derivatives of the left-hand side of equation (1) by the following difference expressions:

$$(u_t)_n^k \cong \frac{u_n^{k+1} - u_n^k}{\tau},\ \text{or}\ (u_t)_n^k \cong \frac{u_n^k - u_n^{k-1}}{\tau},$$

$$(u_{xx})_n^k \cong \frac{u_{n-1}^k - 2u_n^k + u_{n+1}^k}{h^2}. \qquad (4)$$

In accordance with approximation (4) we construct two difference analogues of equation (1) with the unknown net function $v_{h\tau}$:

$$L_h v_{h\tau} = \frac{v_n^{k+1} - v_n^k}{\tau} - \frac{v_{n-1}^k - 2v_n^k + v_{n+1}^k}{h^2} = f_n^k, \qquad (5)$$

$$L_h v_{h\tau} = \frac{v_n^k - v_n^{k-1}}{\tau} - \frac{v_{n-1}^k - 2v_n^k + v_{n+1}^k}{h^2} = f_n^k. \qquad (6)$$

Here f_n^k are the values of a net function $f_{h\tau}$ corresponding to the right-hand side of equation (1), say, $f_n^k = f(x_n,\ t_k)$. We usually assume $f_n^k = f\left(x_n,\ t_k + \frac{\tau}{2}\right)$ for scheme (5) and $f_n^k = f\left(x_n,\ t_k - \frac{\tau}{2}\right)$ for scheme (6).

We can exactly define the initial and the boundary condition for the first boundary-value problem:

$$v_n^0 = u_0(nh)\ (n = 0,\ 1,\ \ldots,\ N),$$

$$v_0^k = \varphi_0(k\tau),\ v_N^k = \psi_0(k\tau),\ (k = 0,\ 1,\ \ldots,\ M). \qquad (7)$$

In the case of the second and third boundary-value problems, the boundary conditions are approximated with the use of formulas similar to relation (4).

Four-point schemes (5) and (6) are shown in Fig. 10.3. Scheme (5) is *explicit* and scheme (6) is *implicit*.

This definition is due to the fact that scheme (5) in the explicit form defines the successive, in time, values of

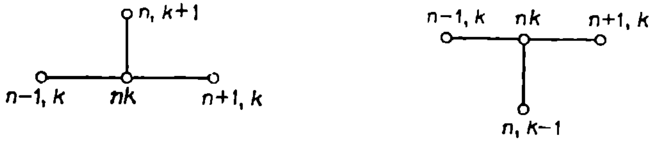


Fig. 10.3

the unknown net function relative to the preceding values. Indeed, setting $r = \tau/h^2$, it is easy to find from relation (5) that

$$v_n^{k+1} = r\left(v_{n-1}^k + v_{n+1}^k\right) + (1-2r)\, v_n^k + \tau f_n^k. \tag{8}$$

Thus, using conditions (7) and the explicit formula (8), we can find, in succession, any value v_n^k. Consequently, a solution of system (7), (8) exists and is unique.

It is different with scheme (6). We rewrite it as follows:

$$r v_{n-1}^k - (1+2r)\, v_n^k + r v_{n+1}^k = -\left(v_n^{k-1} + \tau f_n^k\right). \tag{9}$$

This scheme yields the values of the required net function in an implicit form, i.e. as a system of equations. We can show that a solution of system (7), (9) exists and is unique. It is usually found by the method of factorization which we shall not consider here.

The order of approximation for schemes (5) and (6) can be determined with the use of the appropriate Taylor's formulas in the same way as we did for Poisson's equation in 10.4. As a result we find that the difference schemes (5), (7) and (6), (7) approximate problem (1)-(3) with an error $O(\tau + h^2)$, i.e.

$$\| L_h u_{h\tau} - f_{h\tau} \| \leqslant M(\tau + h^2). \tag{10}$$

The theory of difference schemes proves the validity of the following properties.

1°. *For* $r \leqslant 1/2$, *the explicit scheme* (5), (7) *has a unique solution and is stable, and for* $r \geqslant 1/2$ *it is unstable.*

2°. *The implicit scheme* (6), (7) *has a unique solution and is stable for any* r.

Thus, on the basis of the theorem from 10.4, relation (10) and the properties just formulated, we can assert the convergence of the explicit scheme for $r \leqslant 1/2$ and the implicit scheme for any h and τ with the error $O(\tau + h^2)$.

Example. Solve problem (1)-(3) for $f = 0$, $u_0 = x(1 - x)$, $\varphi_0 = \psi_0 = 0$, $T = 0.1$.

△ In this case equation (1) and conditions (2), (3) assume the form

$$u_t - u_{xx} = 0 \quad (0 < x < 1, \quad 0 < t \leqslant 0.1),$$
$$u(x, 0) = x(1 - x) \quad (0 \leqslant x \leqslant 1),$$
$$u(0, t) = u(1, t) = 0 \quad (0 \leqslant t \leqslant 0.1).$$

We take the explicit scheme (8) as the scheme for computations. We set $h = 0.25$ and then $\tau \leqslant 0.03$. Since $T = 0.1$, we take $\tau = 0.025$ for M to be an integer ($M = 4$). We find that $r = \tau/h^2 = 0.4$. The computational formula has the form

$$v_n^{k+1} = 0.4\left(v_{n-1}^k + v_{n+1}^k\right) + 0.2v_n^k \quad (n = 1, 2, 3, \; k = 0, 1, 2, 3),$$
$$v_n^0 = \frac{n(4-n)}{16}, \; v_0^k = v_4^k = 0 \quad (n = 1, 2, 3, \; k = 1, 2, 3, 4).$$

Thus we get initial conditions $v_1^0 = 0.1875$, $v_2^0 = 0.2500$, $v_3^0 = 0.1875$ and boundary conditions $v_0^k = v_4^k = 0$.

At the first step we have

$$v_1^1 = 0.4\left(v_0^0 + v_2^0\right) + 0.2v_1^0 = 0.1375,$$
$$v_2^1 = 0.4\left(v_1^0 + v_3^0\right) + 0.2v_2^0 = 0.2000.$$

By virtue of symmetry $v_3^1 = v_1^1 = 0.1375$. The calculations at the next steps are similar.

All the calculations are given in Table 10.6. ▲

Table 10.6

n \ k	0	1	2	3	4
v_0^k	0	0	0	0	
v_1^k	0.1875	0.1375	0.1075	0.0815	0.0627
v_2^k	0.2500	0.2000	0.1500	0.1160	0.0884
v_3^k	0.1875	0.1375	0.1075	0.0815	0.0627
v_4^k	0	0	0	0	

10.6. Difference Schemes for Solving the Equation of Oscillation of a String

Let us consider the first boundary-value problem for the equation of the oscillation of a string in a rectangle $\bar{D}$ $\{0 \leqslant x \leqslant 1,\ 0 \leqslant t \leqslant T\}$. We have to find a solution of the problem

$$Lu = u_{tt} - u_{xx} = f \quad (0 < x < 1,\ 0 < t \leqslant T), \tag{1}$$

$$u(x, 0) = u_0(x),\ u_t(x, 0) = u_1(x) \quad (0 \leqslant x \leqslant 1), \tag{2}$$

$$u(0, t) = \varphi_0(t),\ u(1, t) = \psi_0(t) \quad (0 \leqslant t \leqslant T) \tag{3}$$

continuous in $\bar{D}$.

The application of the method of finite differences to the solution of problem (1)-(3) differs but slightly from its application to the equation of heat conduction*. The domain $\bar{D}$ is covered by a net $\bar{\omega}_{h\tau}$. The difference is in the approximation of the second derivative with respect to the variable t:

$$(u_{tt})_n^k \cong \frac{u_n^{k-1} - 2u_n^k + u_n^{k+1}}{\tau^2}. \tag{4}$$

The difference approximation of the equation assumes the form

$$L_h v_{h\tau} \equiv \frac{v_n^{k-1} - 2v_n^k + v_n^{k+1}}{\tau^2} - \frac{v_{n-1}^k - 2v_n^k + v_{n+1}^k}{h^2} = f_n^k. \tag{5}$$

The initial conditions are approximated as follows:

$$v_n^0 = u_0(nh),\quad \frac{v_n^1 - v_n^{-1}}{\tau} = u_1(nh) \quad (n = 0,\ 1,\ \ldots,\ N). \tag{6}$$

The boundary conditions are approximated in the same way as for the equation of heat conduction:

$$v_0^k = \varphi_0(k\tau),\quad v_N^k = \psi_0(k\tau) \quad (k = 0,\ 1,\ \ldots,\ M). \tag{7}$$

Five-point scheme (5) of the "cross" type is shown in Fig. 10.2.

---

* In this section we use designations and concepts accepted in 10.4 and 10.5.

The quantity v_n^{-1} is an apparent (dummy) unknown which can be found from relation (6) and substituted into equation (5). In this case we get a simple explicit scheme ($\gamma = \tau/h$):

$$v_n^{k+1} = -v_n^{k-1} + \gamma^2 \left(v_{n-1}^k + v_{n+1}^k\right) + 2\,(1-\gamma^2)\, v_n^k + \tau^2 f_n^k. \tag{8}$$

The order of approximation of the difference scheme (6)-(8) can be found in the same way as it was done in 10.4 for Laplace's equation. The verification shows that the error of the approximation scheme (6)-(8) is $O(\tau^2 + h^2)$, and in addition, this scheme is stable for $\gamma^2 = (\tau/h)^2 \leqslant 1/(1+\varepsilon)$, $\varepsilon > 0$. Thus it converges with an error of the order of $O(\tau^2 + h^2)$ under the indicated condition.

Example. Solve problem (1)-(3) for $f = 0$, $u_0 = x\,(1-x)$, $u_1 = \varphi_0 = \psi_0 = 0$, $T = 0.6$.

△ We set $h = 0.25$ and then $\tau < 0.25$. Since $T = 0.6$, we take $\tau = 0.2$ for k to be an integer ($M = 3$). We calculate $\gamma^2 = (\tau/h)^2 = 0.64$. The computational formula has the form

$$v_n^{k+1} = -v_n^{k-1} + 0.64\left(v_{n-1}^k + v_{n+1}^k\right) + 0.72 v_n^k \quad (n = 1, 2, 3),$$

$$v_n^0 = \frac{n\,(4-n)}{16} \quad \frac{v_n^1 - v_n^{-1}}{0.2} = 0 \quad (n = 0, 1, 2, 3, 4).$$

$$v_0^k = v_4^k = 0 \quad (k = 0, 1, 2, 3, 4).$$

Thus we get initial conditions $v_1^0 = 0.188$, $v_2^0 = 0.250$, $v_3^0 = 0.188$, $v_n^{-1} = v_n^1$ and boundary conditions $v_0^k = v_4^k = 0$ ($k = 0, 1, 2, 3, 4$).

At the first step we need the value of the apparent quantity v_1^{-1} to calculate v_1^1. We find it from the initial condition $v_1^{-1} = v_1^1$. Thus at the first step we have

$$v_n^1 = -v_n^1 + 0.64(v_{n-1}^0 + v_{n+1}^0) + 0.72 v_n^0.$$

Hence

$$v_1^1 = 0.32\,(v_0^0 + v_2^0) + 0.36 v_1^0 = 0.148,$$

$$v_2^1 = 0.32\,(v_1^0 + v_3^0) + 0.36 v_2^0 = 0.210.$$

By virtue of the symmetry of the problem, $v_3^1 = v_1^1 = 0.148$.

At the second step we obtain

$$v_1^2 = -v_1^0 + 0.64\,(v_0^1 + v_2^1) + 0.72 v_1^1 = 0.053,$$

$$v_2^2 = -v_2^0 + 0.64\,(v_1^1 + v_3^1) + 0.72 v_2^1 = 0.091.$$

By virtue of the symmetry of the problem, $v_3^2 = v_1^2 = 0.053$. By analogy we make calculations at the next steps.

All the calculations are shown in Table 10.7. ▲

Table 10.7

h	0	1	2	3
v_0^k	0	0	0	0
v_1^k	0.188	0.148	0.053	−0.052
v_2^k	0.250	0.210	0.091	−0.077
v_3^k	0.188	0.148	0.053	−0.052
v_4^k	0	0	0	0

Exercises

1. Find an approximate solution of the Laplace equation $\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=0$ for a square under the indicated boundary conditions:

(a)

		16.18	38.63		
0.00	●	●	●	●	50.00
0.00	●	○	○	●	30.10
0.00	●	○	○	●	12.38
0.00	●	●	●	●	4.31
		26.15	29.34		

(b)

		17.98	39.02		
0.00	●	●	●	●	50.00
0.00	●	○	○	●	30.10
0.00	●	○	○	●	12.38
0.00	●	●	●	●	4.31
		29.05	29.63		

2. Find an approximate solution of the Laplace equation $\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=0$ with a stepsize $h=1/6$ for a square under the indicated boundary conditions:

		9.81	19.78	29.12	40.16	42.31	
0.00	×	×	×	×	×	×	×50.00
0.00	×	○	○	○	○	○	×40.16
0.00	×	○	○	○	○	○	×33.11
0.00	×	○	○	○	○	○	×19.14
0.00	×	○	○	○	○	○	×13.00
0.00	×	○	○	○	○	○	×6.98
0.00	×	×	×	×	×	×	×4.31
		17.28	31.96	40.00	30.50	17.28	

3. Find an approximate solution of the equation $\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t}$ which satisfies the conditions $u(x, 0) = g_0(x)$, $u(0, t) = f_0(t)$, $u(1, t) = f_1(t)$ for the values $0 \leqslant t \leqslant T$, taking a step $h = 0.1$ with respect to the argument x. Consider two variants of the boundary conditions:

(a) $g_0(x) = (1.1x^2 + 1.1) \sin \pi x$, $f_0(t) = f_1(t) = 0$, $T = 0.02$, $r = 1/2$.

(b) $g_0(x) = (1.1x^2 + 2.3) e^{-x}$, $f_0(t) = 2{,}3$, $f_1(t) = 3.4e^{-1}$, $= 0.01$, $r = 1/6$.

Answers to Exercises

Chapter 1

1. (a) 2.7546, 2.755, 2.75, 2.8, 3, (b) 3.1416, 3.142, 3.14, 3.1, 3, (c) 0.5645, 0.565, 0.56, 0.6, 1, (d) 4.194, 4.19, 4.2, 4, (e) 0.6065, 0.607, 0.61, 0.6, 1. **2.** (a) 1.14, $\Delta_a = 0.0026$, $\delta_a = 0.23\%$, (b) 0.0102, $\Delta_a = 0.00005$, $\delta_a = 0.5\%$, (c) 0.124, $\Delta_a = 0.0005$, $\delta_a = 0.41\%$, (d) 922, $\Delta_a = 0.45$, $\delta_a = 0.049\%$, (e) 0.00246, $\Delta_a = 0.000002$, $\delta_a = 0.082\%$. **3.** (a) 0.018, (b) 0.099, (c) 0.047, (d) 2.0, (e) 0.00035. **4.** (a) 3, (b) 4, (c) 3, (d) 2, (e) 3. **5.** (a) **The second**, (b) **the second**, (c) **the first**, (d) **the second**, (e) **the second**. **6.** $a = 47.5$. **7.** 46.39. **8.** 3.29. **9.** $\Delta_a = 0.0005$, $\delta_a = 0.0075\%$, (b) $\Delta_a = 0.0005$, $\delta_a = 0.003\%$. **10.** (a) $y = 0.085$, $\Delta_y = 0.0012$, $\delta_y = 1.4\%$, (b) $y = 1.20$, $\Delta_y = 0.056$, $\delta_y = 4.7\%$, (c) $y = 0.0552$, $\Delta_y = 0.00043$, $\delta_y = 0.77\%$, (d) $y = 2.747$, $\Delta_y = 0.0090$, $\delta_y = 0.33$. **11.** (a) $s = 0.594$, (b) $s = 0.687$.

Chapter 2

1. (a) $\begin{bmatrix} 6 & 2 & 4 \\ 9 & 9 & 6 \\ 8 & 9 & 1 \end{bmatrix}$, (b) $\begin{bmatrix} 1 & 5 & -5 \\ 3 & 10 & 0 \\ 2 & 9 & -7 \end{bmatrix}$.

2. (a) $\begin{bmatrix} 8 & -56 & 54 \\ -30 & -100 & 146 \\ 118 & -82 & 28 \end{bmatrix}$, (b) $\begin{bmatrix} -72 & -72 & 78 \\ 36 & 54 & -6 \\ 66 & 240 & 88 \end{bmatrix}$.

3. (a) $\begin{bmatrix} 5 & 10 & -10 & 15 \\ 7 & 14 & -14 & 21 \\ -3 & -6 & 6 & -9 \\ 2 & 4 & -4 & 6 \end{bmatrix}$, (b) $\begin{bmatrix} 4 & 5 \\ 8 & 10 \\ 12 & 15 \end{bmatrix}$, (c) 409.

4. (a) $\begin{bmatrix} -9 \\ 20 \\ -18 \end{bmatrix}$, (b) $\begin{bmatrix} 8 \\ 2 \\ 2 \end{bmatrix}$. **5.** (a) 22, (b) -26, (c) 4279.1.

6. (a) $A^{-1} = \begin{bmatrix} -12 & -1 & 8 \\ -9 & -1 & 6 \\ -7 & -1 & 5 \end{bmatrix}$, (b) $A^{-1} = \frac{1}{48}\begin{bmatrix} -4 & 10 & -46 & 24 \\ 10 & -1 & 67 & -36 \\ -14 & -13 & 7 & 12 \\ 0 & 0 & -96 & 48 \end{bmatrix}$,

(c) $A^{-1}=\frac{1}{24}\begin{bmatrix} 24 & 0 & 0 & 0 \\ 0 & 12 & 0 & 0 \\ -12 & 9 & 6 & 0 \\ -28 & -5 & 2 & 8 \end{bmatrix}$.

7. $AB=\begin{bmatrix} 24 & 57 & 15 & 31 \\ -4 & 16 & 6 & 11 \\ 16 & 27 & 7 & 14 \\ 10 & 53 & 13 & 29 \end{bmatrix}$. **8.** (a) $A^{-1}=\frac{1}{18}\begin{bmatrix} 1 & 2 & 3 & 2 \\ 2 & -1 & 2 & -3 \\ 3 & -2 & -1 & 2 \\ -2 & -3 & 2 & 1 \end{bmatrix}$,

(b) $A^{-1}=\frac{1}{18}\begin{bmatrix} 6 & -4 & 0 & 8 \\ -12 & 5 & 9 & -1 \\ -12 & 11 & 9 & -13 \\ -6 & -5 & 9 & 1 \end{bmatrix}$. **9.** (a) $A=T_1T_2=$

$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & -5 & 0 & 0 \\ 3 & -4 & -18/5 & 0 \\ 2 & -7 & 36/5 & 18 \end{bmatrix}\cdot\begin{bmatrix} 1 & 2 & 3 & -2 \\ 0 & 1 & 8/5 & -1/5 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 1 \end{bmatrix}$, $A^{-1}=T_2^{-1}T_1^{-1}=$

$\begin{bmatrix} 1 & -2 & 1/5 & 2 \\ 0 & 1 & 8/5 & -3 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 & 0 \\ 2/5 & -1/5 & 0 & 0 \\ 7/18 & 4/18 & -5/18 & 0 \\ -2/18 & -3/18 & 2/18 & 1/18 \end{bmatrix}=$

$\frac{1}{18}\begin{bmatrix} 1 & 2 & 3 & 2 \\ 2 & -1 & 2 & -3 \\ 3 & -2 & -1 & 2 \\ -2 & -3 & 2 & 1 \end{bmatrix}$, (b) $A=R_1R_2=\begin{bmatrix} 3 & 0 & 0 & 0 \\ 2 & 7/3 & 0 & 0 \\ 3 & 1 & 1/7 & 0 \\ 1 & 8/3 & -9/7 & 18 \end{bmatrix}\times$

$\begin{bmatrix} 1 & -2/3 & 2/3 & 0 \\ 0 & 1 & -1/7 & -6/7 \\ 0 & 0 & 1 & 13 \\ 0 & 0 & 0 & 1 \end{bmatrix}$, $A^{-1}=R_2^{-1}=R_1^{-1}=$

$\begin{bmatrix} 1 & 2/3 & -4/7 & 8 \\ 0 & 1 & 1/7 & -1 \\ 0 & 0 & 1 & -13 \\ 0 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1/3 & 0 & 0 & 0 \\ -2/7 & 3/7 & 0 & 0 \\ -5 & -3 & 7 & 0 \\ -1/3 & -5/18 & 9/18 & 1/18 \end{bmatrix}=$

$\frac{1}{18}\begin{bmatrix} 6 & -4 & 0 & 8 \\ -12 & 5 & 9 & -1 \\ -12 & 11 & 9 & -13 \\ -6 & -5 & 9 & 1 \end{bmatrix}$, **10.** (a) $X=\begin{bmatrix} 3 & 2 & -1 \\ 0 & -1 & 2 \\ 5 & 7 & 1 \end{bmatrix}$,

(b) $X=\begin{bmatrix} 0 & -1 & 2 \\ 1 & 0 & -2 \\ 3 & 1 & 2 \end{bmatrix}$, **11.** (a) $r=2$, (b) $r=2$. **12.** It is linearly dependent, (b) it is linearly independent. 13.. The basis is composed, say, of the vectors $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_4$; $\mathbf{x}_3 = \mathbf{x}_1 - \mathbf{x}_2$. 14. $\mathbf{y} = (5/4,\ 1/4;\ -1/4,\ -1/4)$.

Chapter 3

1. (a) $x_1 = -(11/7)x_3$, $x_2 = -(1/7)x_3$, (b) $x_1 = (3x_3 - 13x_4)/17$, $x_2 = (19x_3 - 20x_4)/17$. 2. The general solution is $x_1 = (x_3 - 9x_4 - 2)/11$, $x_2 = (-5x_3 + x_4 + 10)/11$; the particular solution is $x_1 = -1$, $x_2 = 1$, $x_3 = 0$, $x_4 = 1$, (b) the general solution is $x_3 = 22x_1 - 33x_2 - 11$, $x_4 = -16x_1 + 24x_2 + 8$, the particular solution is $x_1 = 1$, $x_2 = 0$, $x_3 = 11$, $x_4 = -8$. 3. (a) $x_1 = 0$, $x_2 = -1$, $x_3 = 2$, (b) $x = 2$, $y = -2$, $z = 1$. 4. (a) $x_1 = 1$, $x_2 = -1$, $x_3 = 1$, $x_4 = -1$, (b) $x_1 = -1$, $x_2 = 2$, $x_3 = 0$, $x_4 = 3$. 5. (a) $x_1 = 1.120$, $x_2 = -0.341$, $x_3 = -0.008$, (b) $x = 0.008$, $y = -0.231$, $z = 0.042$. 6. (a) $d = 88$, (b) $d = 2111.97$.

7. (a) $A^{-1} = \begin{bmatrix} 10/3 & -7/6 & 1/2 & -1/6 \\ -5/3 & 5/6 & -1/2 & -7/6 \\ 1 & -1/2 & 1/2 & 3/2 \\ 1 & -1/2 & 1/2 & 1/2 \end{bmatrix}$,

(b) $A^{-1} = \begin{bmatrix} 1.19 & -0.31 & -0.82 & -0.12 \\ -0.17 & 1.57 & 1.23 & 0.70 \\ -1.75 & 0.11 & 0.30 & 0.87 \\ -0.12 & -2.92 & -1.09 & 0.17 \end{bmatrix}$.

8. (a) $x_1 = -0.72$, $x_2 = 1.88$, $x_3 = -0.92$, $x_4 = -1.94$, (b) $x = 1.22$, $y = -067$, $z = 0.35$. 10. (a) $\| A \|_1 = 1.9$, $\| A \|_2 = 1.9$, $\| A \|_3 = 2.55$, (b) $\| A \|_1 = 1.45$, $\| A \|_2 = 1.07$, $\| A \|_3 = 1.20$. 11. (a) $x_1 = 1$, $x_2 = -1$, $x_3 = 2$, $x_4 = 0$, (b) $x_1 = 1/2$, $x_2 = 3/2$, $x_3 = -1/2$, $x_4 = -2$, 12. (a) $x_1 = -2$, $x_2 = 2$, $x_3 = -3$, $x_4 = 3$, (b) $x_1 = 1$, $x_2 = 1$, $x_3 = -1$.

Chapter 4

1. The remainder is $r = P_5(3) = 430$, the quotient is $P_4^{(1)} = x^4 + 6x^3 + 16x^2 + 48x + 143$. 2. Yes, it is. 3. (a) 0.423, (b) 0.940, (c) 1.386, (d) 1.221, (e) 0.809, (f) 0.309. 4. (a) 3.464, (b) 7.483, (c) 6.481.

Chapter 5

1. (a) — 1.325, (b) 1.180, (c) — 1.876, 0.578. (d) 0.781, 2.401, (e) 0.0, 0.399, 6.352, (f) 0.310, 4.0. 2. (a) 1.213, (b) 0.706, (c) 0.841, (d) —0.438, 0.438, (e) 0.0, 0.787, (f) 1.897. 3. (a) —4.071, 0.468, 0.993, (b) —0.695, —3.067, 3.757, (c) —3.523, —1.567, 1.086, (d) 0.398, 4.862, (e) 0.0, 2.753, (f) 0.739. 4. (a) 0.760, (b) — 2.258, (c) —0.465, (d) —0.567, —0.335, 0.0, (e) 3.473, (f) 1.422. 5. (a) —0.532, 0.653, 2.879, (b) —0.475, 1.395, (c) —1.582, 0.402, 1.373, (d) —1.453, 1.164. 6. (a) 0.187, (b) 0.755, (c) 0.739, (d) 0.607, (e) 0.672.

Chapter 6

1. (a) $\lambda^3-2\lambda^2-66\lambda+1$, (b) $\lambda^4-4\lambda^3-6\lambda-41$, (c) $\lambda^4-4\lambda^3+16\lambda^2-16$. **2.** (a) $\lambda_1=7$, $x_1=c\begin{bmatrix}1\\1\end{bmatrix}$, $\lambda_2=-2$, $\mathbf{x}_2=c\begin{bmatrix}4\\-5\end{bmatrix}$, (b) $\lambda_1=-2$, $\mathbf{x}_1=c_1\begin{bmatrix}-2\\1\\0\end{bmatrix}+c_2\begin{bmatrix}1\\0\\1\end{bmatrix}$. **3.** (a) $\lambda^3+\lambda^2-3\lambda-7$, (b) $\lambda^4+8\lambda^3+6\lambda^2+6\lambda-54$, (c) $\lambda^4+\lambda^3-6x^2-18\lambda$. **4.** (a) $\lambda^4-7\lambda^3+15\lambda^2-2\lambda-34$, (b) $\lambda^4+\lambda^3+7\lambda^2-20\lambda-54$. **6.** $D(\lambda)=\lambda^3-3\lambda^2+3\lambda-1$, $\lambda_1=\lambda_2=\lambda_3=1$, $\mathbf{x}^{(1)}=\mathbf{x}^{(2)}=\mathbf{x}^{(3)}=\begin{bmatrix}1\\1\\-1\end{bmatrix}$, if $=\begin{bmatrix}1\\0\\0\end{bmatrix}$. **7.** $\lambda_1\cong 4.46$, $\lambda_2=1.59$.

Chapter 7

1. 20.819, **2.** 24.680. **3.** $L_3(x)=-(1/15)x^3-(3/20)x^2+(241/60)x-3.9$. **4.** $L_2(x)=0.0735x^2-0.4530x+0.7474$, $\Delta_1=0.23\times10^{-4}$. **7.** 3.37215. **8.** 4.379. **9.** 1.43612. **10.** 0.75487. **13.** (a) 0.6115 ± 0.00013, (b) 0.9409 ± 0.0007, (c) 0.9456 ± 0.002, (d) 0.8007 ± 0.0009, (e) 0.3156 ± 0.00012, (f) 2.4505 ± 0.0018. **14.** $Q_2(x)=2-0.7\cos 2x-0.3\sin 2x-0.3\cos 4x+b_2\sin 4x$. **15.** $Q_2(x)=-1+(7/3)\cos 2\pi x-(2/\sqrt{3})\sin 2\pi x-\cos 4\pi x$.

Chapter 8

1. 15.160. **2.** 1.429. **3.** 2.28. **4.** 2.002. **5.** 0.107250. **6.** 0.67363. **7.** 0.6931472. **8.** 0.007. **9.** 0.0087. **10.** (a) 0.239, (b) 0.223, (c) 1.000, (d) 0.8349, (e) 1.4627, (f) 1.5625, (g) 1.333, (h) 0.460, (i) 1.718. **11.** (a) 0.754 ± 0.002 (for $h_0=3°$), (b) -0.471 ± 0.002 (for $h_0=7°$), (c) 2.421 ± 0.007 (for $h_0=1°$), (d) 0.953 ± 0.002 (for $h_0=6°$), (e) 0.892 ± 0.0015 (for $h_0=3°$), (c) 1.438 ± 0.003 (for $h_0=3°$).

Chapter 9

1. (a) $y_3+8x^2+(56/3)x^3+18x^4+8x^5+(4/3)x^6$, (b) $y_3=1-x+x^2-(1/3)x^3+(1/24)x^4$. **2.** $y(x)=1+2x-0.7x^2-0.2567x^3+0.051x^4+0.00147x^5-0.00101x^6$. **3.** $y=x^3/3+x^7/63+\ldots$ **4.** (a) $y_1=-1.1$, $y_2=-1.18$, $y_3=-1.238$, $y_4=-1.2718$, $y_5=-1.2790$, (b) $y_1=0$, $y_2=0{,}01$, $y_3=0.0278$, $y_4=0.0524$, $y_5=0.08192$, $y_6=0.115536$, $y_7=0.152429$, $y_8=0.191943$, $y_9=0.233554$, $y_{10}=0.276844$. **5.** $y_0=1$, $y_1=1.1836$, $y_2=1.3426$, $y_3=1.4850$, $y_4=1.6152$, $y_5=1.7362$. **6.** $y_0=1$, $y_1=1.1867$, $y_2=1{,}3484$, $y_3=1.4938$, $y_4=1.6272$, $y_5=1.7542$. **7.** (a) $y_0=0$, $y_1=0.10536$, $y_2=0.223136$, $y_3=0.356601$, $y_4=0.510424$, $y_5=0.691497$, $y_6=0.910454$, $y_7=1.184648$, $y_8=1.544491$, $y_9=2.048721$, $y_{10}=2.827617$, (b) $y_0=$

2.00, $y_1 = 1.81$, $y_2 = 1.64$, $y_3 = 1.49$, $y_4 = 1.36$, $y_5 = 1.25$, $y_6 = 1.16$, $y_7 = 1.09$, $y_8 = 1.04$, $y_9 = 1.01$, 8. $y_4 = 0.8110$, $y_5 = 0.8196$, $y_6 = 0.8464$, $y_7 = 0.8898$, $y_8 = 0.9480$, $y_9 = 1.0197$, $y_{10} = 1.1037$.

Chapter 10

1. (a)

0.00	16.18	38.63	50.00
0.00	14.12	26.09	30.10
0.00	15.20	20.53	12.38
0.00	26.15	29.34	4.31

(b)

0.00	17.88	39.92	50.00
0.00	15.18	36.39	30.10
0.00	16.37	21.26	12.38
0.00	29.05	29.63	4.31

2.

	9.81	19.78	29.12	40.16	42.31	
0.00	8.97	17.58	25.36	32.18	36.11	40.16
0.00	8.68	16.00	22.29	26.86	29.69	33.11
0.00	8.36	15.59	20.71	23.05	22.62	19.11
0.00	9.43	17.22	21.71	21.85	18.55	13.00
0.00	12.20	22.09	26.96	24.01	16.70	6.98
	17.28	31.96	40.00	30.50	17.28	

3.

	t	1	2	3	4	5	6	7	8	9
(a)	0.000	0.343	0.672	0.970	1.213	1.375	1.423	1.327	1.062	0.618
	0.005	0.336	0.656	0.943	1.172	1.318	1.351	1.243	0.973	0.531
	0.010	0.328	0.639	0.914	1.131	1.262	1.281	1.162	0.887	0.486
	0.015	0.320	0.621	0.885	1.088	1.206	1.212	1.084	0.824	0.443
	0.020	0.311	0.602	0.855	1.045	1.150	1.145	1.018	0.764	0.412
(b)	0.000	2.091	1.919	1.777	1.660	1.562	1.480	1.410	1.350	1.297
	0.017	2.097	1.924	1.781	1.663	1.564	8.482	1.411	1.351	1.298
	0.033	2.102	1.929	1.785	1.666	1.567	1.484	1.413	1.352	1.299
	0.050	2.106	1.934	1.789	1.670	1.570	1.486	1.415	1.354	1.300
	0.067	2.110	1.939	1.794	1.673	1.572	1.488	1.416	1.355	1.301

Index

Absolute, error 17
  value of a matrix 89
    of a vector 90
Adams' extrapolation method 452, 454, 457
Adjoined matrix 58
Aitken's iterated interpolation 339
Algebraic, adjunct 50
  equation 181
  function 180
  polynomial 165
Analytic method of representing a function 283
Approximate, number 14
  value of a number 17
Arithmetic progression 347
Augmented matrix of a system of equations 110

Back substitution 127, 130
Base, equation 115
  minor 99
  unknown 115
Basic elimination procedure 121
Bezout's theorem 168
Bisection method 194
Bordered matrix 79
Bordering 79
Boundary conditions 475
Boundary-value problem 465
  classical solution of 476
  homogeneous 467
  linear 467

Cauchy's problem 425, 474
  for hyperbolic equations 477
  for parabolic equations 478
  for an unbounded domain 476
Central difference 309, 311
Characteristic, determinant 239
  equation 239
  polynomial 239
  value 237
Chebyshev's method "of fitting" 292
  polynomial 341, 342
Cholesky's method 144
Classical solution of a boundary-value problem 476
Coefficient, of Cotes, 400
  of an equation 181
  of Fourier 351
  of a system of equations 108
Cofactor 50
Columns, matrix 40
  vector 40
Conjugate matrix 58
Consistent system of equations 109
Constant, term 108, 165
  unknown 115
Control sum 130
Convergent, method 415
  sequence 348
  series 349
Coordinates of a vector 41, 99
Cramer's rule 112

Deficiency of a matrix 91
Determinant 48
  principal diagonal of 48
  properties of 52-55
  of the second order 48
  secondary diagonal of 48
  of the third order 49

Determinate system of equations 109
Diagonal, matrix 41
  minor 67
Difference of matrices 43
Difference, method 483
  problem 489
Differential equation 423
    degree of 434
    elliptic 472
    hyperbolic 472
    integral of 424
      general 424
    ordinary 423
      of order n 424
    parabolic 472
    partial 423
      solution of 424
Dimensions, of a matrix 40
  of space 98
Dirichlet's problem 479
Discriminant 472
Divergent, sequence 348
  series 349
Divided difference 329
  of order zero 329
Domain, of convergence of a functional series 350
  of definition of a function 190
  of existence of a function 190
  of permissible values of an equation 179
Doubtful digit 21

Eigenvalue 237
Eigenvector 237
Elementary transformation, of a determinant 55
    of a matrix 93
    of a system of equations 109
Elements of a vector 41
Equal matrices 42
Equation, of heat conduction 474
  of Laplace 478
  of oscillation of a string 474
  of Poisson 474
Equispaced interpolation nodal points 307
Equivalent, matrices 93
  systems of equations 109
Error 17
  of approximation 15
    of a boundary condition 488
    of an equation 487
  of interpolation 295, 297
  of a method 15
  of rounding 15
Euler's, broken line 436
  method 434
Exact method of solving a problem 240
Expansion, of a determinant according to the elements of a row (column) 51
  of a function in a Fourier series 351
Extrapolation 293

Finite difference 289, 309
    backward 309, 311
    central 309, 311
    forward 309, 311
Finite-dimensional space 98
Formula, of Adams 454
  of Gauss 410
  rectangular 382, 384
  of Simpson 402
    compounded 418
Forward substitution 127, 130
Fourier, analysis 357
  coefficient 351
  series 351
Free term of a system of equations 108
Frobenius matrix 261
Function 283
Functional dependence 283
  sequence 348
  series 349
    trigonometric 351
Fundamental, system of equations 118
  theorem of algebra 224

Gauss' scheme 127
Gaussian, elimination 124
  quadrature formula 410

General, integral of a differential equation 424
  solution of a differential equation 424
Generalized Horner's scheme 170
Geometric progression 347
Graph of a function 284
Graphical method, of representing a function 284
  of separating roots 184

Harmonic, analysis 347
  function 480
Hermite polynomial 345
Homogeneous, boundary-value problem 467
  system of equations 109
  Horner's, method 231
  scheme 167
Hypermatrix 78

Identity matrix 41
Implicit scheme 496
Incompatible system of equations 109
Inconsistent system of equations 109
Indeterminate system of equations 109
Infinite-dimensional space 98
Initial, conditions 425, 474
  data 425
  interval 454
  value 425
Initial-value problem 425
Input data 481, 482, 489
Integral, curve 424
  rational function 180
Interpolation 293
  backward 366
  error of 295
  formula 398
  linear 297
  in narrow sense 293
  by polynomials 294
  quadratic 297
  in the sense of Hermite 345
  in the sense of Lagrange 345
  in tables 336
Interpolation polynomial 295
  Bessel's 322, 323
  Lagrange's 304, 306
  Newton's 325, 326
  Stirling's 320, 322
Inverse matrix 57
Inversion of a matrix 57
  by expanding it in a product of two triangular matrices 72
  by partitioning it into blocks 82
  by successive bordering 85
Invertible matrix 57
Irrational function 181
Iterative, method of solving a problem 241
  process 148

Kronecker-Capelli theorem 110
Kronecker delta 139

Lagrange's multiplier 304
λ-difference 279
Limit of a sequence 348
Linear, boundary-value problem 467
  combination of vectors 96
  dependence of vectors 96
  independence of vectors 96
  interpolation 297
  space 95
Linearly independent equations 115
Lipschitz condition 425
Loss of accuracy 26
Lower triangular matrix 66

Mantissa of a number 20
Matrix 40
  equation 63-66
  of Frobenius 261
  square 41
  of a system of equations 109
  of transformation of an old basis into a new one 101
Method, of Adams 452, 454, 457
  of Aitken 339
  of bisection 194
  of bordering 92

  of chords 197
  of Danilevsky 260, 263, 270-274
  of direct expansion 241
  of Euler 434
    modifications of 439
  of Euler-Cauchy 441
  of finite differences 483
  of Horner 231
  of iterations 148-150, 176, 214, 277
  of separating roots 184
    analytical 189
    graphical 184
  of Krylov 245, 253
  of Lagrange, of finding the upper bound of positive roots of an equation 229
  of least squares 294
  of Leverrier-Faddeev 254, 258
  of Milne 459
  of Newton, of finding the upper bound of positive roots of an equation 230
  of partitioning into blocks 82
  of Pickard 426
  of representing a function 284
    analytical 282
    graphical 284
    tabular 284
  of Seïdel 156-159
  of successive, differentiation 430
    approximation 148-150, 176 214
    bordering 85
    elimination of unknowns 124
  of tangents 203
  of undetermined coefficients 432
Methods of representing a functional relation 283
  analytical 283
  graphical 284
  tabular 284
Minor, of an element of a determinant 50
  of order *k* of a matrix 91
Mixed boundary-value problem 467
Monotonic function 190
n-dimensional linear space 98
Natural scale 347
Net 484
  domain 484
  function 485
Neumann's problem 479
Newton's, interpolation formula 327
  method of approximation 203
Newton-Cotes quadrature formulas 400
Nodal point 293, 484
Node 485
Noninvertible matrix 57
Nonremovable error 15
Nonsingular matrix 57
Norm, of a matrix 90
    on the set of net functions 487
    of a vector 90
Normal form of a system of linear equations 149
Normalized form of notation of a number 20
Null matrix 42
Numerical, differentiation 374-378
  integration 379-382

Order, of approximation of a boundary-value problem 488
  of convergence of a difference scheme 490
  of correctness of a table of finite differences 315
  of a differential equation 421
  of a number 20
  of a square matrix 41

Partial sum of a series 349
Particular solution of a differential equation 424
Periodic function 351
Piecewise-continuous function 352
Piecewise-monotonic function 352
Pivot, coefffcient 124
  element 121

  - row 121
  - unknown 124
- Power series 349
- Principal diagonal 41
- Principle, of the maximum 480
  - of Runge 455
- Problem, Cauchy's 425, 474
    - for hyperbolic equations 477
    - for parabolic equations 478
    - for an unbouded domain 476
  - Dirichlet's 479
  - with initial conditions 425
  - without initial conditions 475
  - Neumann's 479
- Product, of a matrix by a number 43
  - of two matrices 44
  - of a vector by a scalar 46

- Quadratic interpolation 297
- Quadrature 379
  - formula 372
    - for interpolation 398
    - of Newton-Cotes 396

- Rank of a matrix 91
- Rational function 180
  - entire 180
  - integral 180
- Rational fractional function 181
- Rectangular matrix 40
- Relative error 18
- Root, of an equation 179
  - of multiplicity s 166
  - of a polynomial 166
- Rounding error 15
- Rounding off 16
- Row, matrix 40
  - vector 40
- Rule, of annulus 228
  - of calculating digits 36-37
  - of Descartes 225
  - of rounding off 16
  - of triangles 49
- Runge's principle 455

- Scheme of unique division 127
- Seidel's method 156
- Separated root of an equation 184
- Sequence 347
  - constant 347
  - convergent 348
  - divergent 348
  - general term of 347
- Series 349
  - convergent 349
  - divergent 349
  - of Fourier 351
  - sum of 349
- Significant digit 17
    - valid 20
- Simple root 166
- Singular matrix 57
- Solution, of a differential equation 424
  - of an equation 179
  - of a system of equations 108, 180
- Square matrix 41
- Space basis 98
- Stability of a difference problem 489
- Step of iteration 434
- Step matrix 78
- Subdiagonal matrix 66
- Sum, of matrices 42
  - of series 349
  - of vectors 46
- Superdiagonal matrix 66
- Symmetric matrix 42
- System of equations 108, 114, 118, 124, 179

- Tabular method of representing a function 284
- Theorem, of Bezout 168
  - on a base minor 100
  - on convergence of a difference scheme 490
  - of Dirichlet 352
  - fundamental, of algebra 224
  - of Hamilton-Cayley 245
  - of Kronecker-Capelli 110
  - on linear dependence of vectors 96
  - of Sturm 228
- Theorems, on continuous functions 189

on convergence of interpolation process 215
on determinants 50-52
on quadrature formulas 408-409
Trace of a matrix 242
Transcendental function 181
Transpose of matrix 46
Trial and error method 192
Triangular matrix 66
lower 66
upper 66

Unit, matrix 41
vector 98
Unknowns 108
Unremovable error 15
Upper triangular matrix 66

Valid digit, in a product 30
in a quotient 34
significant digit 20
Value of a decimal position 16
Vector 41
Vector space 95

Zero of a polynomial 166
Zero matrix 42